University of Albert

Automated and Risk-based Box Jenkins Forecasting Model
For Infrastructure System Capital Planning (ISCP)

By

Yong Lu ©

A thesis submitted to the Faculty of Graduate Studies and Research in partial fulfillment of the requirements for the degree of *Master of Science*

In

Construction Engineering and Management

Department of Civil and Environmental Engineering

Edmonton, Alberta

Fall 2004

Library and
Archives Canada

Published Heritage
Branch

395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque et
Archives Canada

Direction du
Patrimoine de l'édition

395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence
ISBN: 0-612-95807-8
Our file Notre référence
ISBN: 0-612-95807-8

The author has granted a non-exclusive license allowing the Library and Archives Canada to reproduce, loan, distribute or sell copies of this thesis in microform, paper or electronic formats.

The author retains ownership of the copyright in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

L'auteur a accordé une licence non exclusive permettant à la Bibliothèque et Archives Canada de reproduire, prêter, distribuer ou vendre des copies de cette thèse sous la forme de microfiche/film, de reproduction sur papier ou sur format électronique.

L'auteur conserve la propriété du droit d'auteur qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou aturement reproduits sans son autorisation.

In compliance with the Canadian Privacy Act some supporting forms may have been removed from this thesis.

While these forms may be included in the document page count, their removal does not represent any loss of content from the thesis.

Conformément à la loi canadienne sur la protection de la vie privée, quelques formulaires secondaires ont été enlevés de cette thèse.

Bien que ces formulaires aient inclus dans la pagination, il n'y aura aucun contenu manquant.

Canada

ACKNOWLEDGEMENT

I would like to thank my advisor Dr. Simaan M. AbouRizk for his supervision of my master research. Without his guidance and financial support, I would never be able to accomplish this thesis. My appreciation is extended to Dr. Mohamed Al-Hussein and Dr. Peter Flynn for serving as members of my examining committee.

Many thanks to all members of the Construction Engineering and Management Group, University of Alberta for their support and suggestions throughout my graduate program. The thesis editing assistance I received from Cam Fraser, and the academic guidance from Dr. Yasser Mohamed and Dr. Lingguang Song made possible the success of this thesis. Most of all, the encouragement from everyone was much appreciated.

I would like to express my appreciation to Mr. Bart Becker from Stantec Consulting Inc. for his assistance and informative feedback during the course of the case study.

I woulld like to recognize my mother Shuqin Qin and elder brother, Qiang Lu, for their support and encouragement from China throughout my life in Canada. Finally, I wish to thank my wife Beibei for her patience, dedication and endless love during my graduate study in the University of Alberta.

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION AND LITERATURE REVIEW 1

1.1 INTRODUCTION 1

1.2 INFRASTRUCTURE SYSTEM CAPITAL PLANNING (ISCP) 2

1.2.1 Concepts *2*

1.2.2 Practices and Problems *3*

1.3 REVIEW OF THE STATE-OF-THE-ART FORECASTING METHODOLOGIES FOR ISCP 5

1.3.1 Introduction *5*

1.3.2 Current Methodologies *6*

1.3.2.1 General Summary 6

1.3.2.2 Time Series Techniques 7

1.3.3 Methodology Study *11*

1.3.4 Limitations of Current Methodologies *13*

1.3.4.1 Theoretical Limitations 14

1.3.4.2 Practical Limitations 15

1.4 RESEARCH OBJECTIVES AND METHODOLOGY 16

1.5 CONCLUSION 17

CHAPTER 2: AUTOMATED BOX JENKINS FORECASTING MODELING 19

2.1 REVIEW OF BOX JENKINS METHODOLOGIES 19

2.2 RATIONALE FOR SELECTING BOX JENKINS METHODOLOGIES 21

2.3 IMPLEMENTATION PROBLEMS ASSOCIATED WITH BOX JENKINS FORECASTING MODELING 23

2.4 AUTOMATING BOX JENKINS FORECASTING MODELING 24

2.5 IMPLEMENTATION OF AUTOMATED BOX JENKINS FORECASTING MODELING 28

2.5.1 Core Algorithms Design and Programs Development *28*

2.5.2 Program Integration – Development of Automated Modeling Tool *30*

2.6 CONCLUSION 42

CHAPTER 3: RISK-BASED BOX JENKINS FORECASTING **43**

3.1 PROBLEMS WITH TRADITIONAL TIME SERIES FORECASTING 43

3.2 RISK-BASED BOX JENKINS FORECASTING APPROACH 43

3.2.1 Simulation Technologies in General *44*

3.2.2 Concerns about Simulation Input Modeling *47*

3.2.3 Box Jenkins-based Monte Carlo Simulation Procedure *51*

3.2.4 Box Jenkins-based Monte Carlo Simulation Methodologies *53*

3.2.5 Box Jenkins-based Monte Carlo Simulation Execution and Output Analysis . *61*

3.3 RISK-BASED BOX JENKINS FORECASTING IMPLEMENTATION 65

3.4 CONCLUSION 67

CHAPTER 4: MODEL VALIDATION AND CASE STUDY **68**

4.1 MODEL VALIDATION 68

4.1.1 Introduction *68*

4.1.2 Validation of Automated Box Jenkins Forecasting *70*

4.1.3 Validation of Risk-based Box Jenkins Forecasting *76*

4.1.3.1 Test on Risk and Uncertainty Analysis Functionality 76

4.1.3.2 Test on Dependent Simulation Input Modeling 82

4.2 CASE STUDY 87

4.2.1 Introduction *87*

4.2.2 Problem Identification and Tentative Solution *88*

4.2.3 Implementation and Results *91*

4.3 CONCLUSION 97

CHAPTER 5: CONCLUSIONS **99**

5.1 SUMMARY OF RESEARCH 99

5.2 RESEARCH CONTRIBUTIONS 101

5.3 LIMITATIONS OF RESEARCH AND RELATED FUTURE WORK 102

REFERENCES AND BIBLIOGRAPHY **104**

APPENDIX A: VALIDATION STUDY FOR BOX JENKINS FORECASTING METHODOLOGIES 108

APPENDIX B: BOX JENKINS-BASED MONTE CARLO SIMULATION VALIDATION 119

APPENDIX C: DEVELOPED COMPUTING CODES FOR BOX JENKINS FORECASTING AND BOX JENKINS-BASED MONTE CARLO SIMULATION 122

APPENDIX D: BOX JENKINS FORECASTING SPS TEMPLATE: USER MANUAL 133

LIST OF TABLES

TABLE 1-1: TIME-SERIES METHODS 8

TABLE 2-1: CHILD ELEMENTS AND CORRESPONDING FUNCTIONALITY 39

TABLE 4-1: TABULATED VALIDATION DATA FROM US CENSUS BUREAU 70

TABLE 4-2: COMPARISON OF FOUR FORECASTING METHODS IN TERMS OF ACCURACY.... 75

TABLE 4-3: DESCRIPTIVE STATISTICS FOR FOUR EXAMPLE TIME POINTS........ 79

TABLE 4-4: RANK AND PERCENT ANALYSIS 81

TABLE 4-5: COMPARISON BETWEEN TRADITIONAL FORECASTING AND BOX JENKINS SIMULATION 82

TABLE 4-6: DESCRIPTIVE STATISTICAL ANALYSIS (TRADITIONAL METHODOLOGIES) 86

TABLE 4-7: INFLATION RATES FOR THREE COST ITEMS (ORIGINAL ANALYSIS)........ 89

TABLE 4-8: SELECTION OF COST INDEX FOR CORRESPONDING COST INFLATION ITEM...... 90

TABLE 4-9: COMPARISON OF INFLATION RATE ESTIMATION 93

TABLE 4-10: COMPARISON OF ANNUAL CASH FLOW ESTIMATION 94

TABLE A-1: TABLE FOR THE SELECTION OF *K* VALUE FOR MOVING AVERAGE METHOD ... 110

TABLE A-2: SUMMARY OF TREND REGRESSION ANALYSIS 112

TABLE B-1: SIMULATED OBSERVATIONS FOR TESTING PERIOD OF TIME (ITERATION 1-50) 120

TABLE B-2: SIMULATED OBSERVATIONS FOR TESTING PERIOD OF TIME (ITERATION 51-100) 121

LIST OF FIGURES

FIGURE 2-1: FLOW CHART OF BOX JENKINS FORECASTING MODELING ALGORITHMS 31

FIGURE 3-1: SCHEMATIC ILLUSTRATION OF MONTE CARLO SIMULATION PROCESS – "CAPITAL INVESTMENT ANALYSIS FOR ENGINEERING AND MANAGEMENT", CANADA (1996) 46

FIGURE 3-2: TAXONOMY OF INPUT MODELS (LEEMIS 2001) 50

FIGURE 3-3: BOX JENKINS-BASED MONTE CARLO SIMULATION ALGORITHMS........ 62

FIGURE 4-1: TIME-SERIES PLOT OF VALIDATION DATA FROM US CENSUS BUREAU 70

FIGURE 4-2: COMPARISON OF FOUR FORECASTING METHODS IN TERMS OF ACCURACY... 74

FIGURE 4-3: TIME-SERIES PLOT FOR 100 SIMULATION PATHS........ 78

FIGURE 4-4: EMPIRICAL CDF AND HISTOGRAM FOR FOUR SAMPLE TIME POINTS 80

FIGURE 4-5: SCATTERPLOT OF CORRELATION BETWEEN OBSERVATIONS WITH DIFFERENT LAG (1, 2, AND 3)........ 84

FIGURE 4-6: TIME-SERIES PLOT FOR TRADITIONAL INPUT MODELING METHOD 85

FIGURE 4-7: SIMULATED RANDOM VARIATES FOR CONSTRUCTION COST INDEX.............. 91

FIGURE 4-8: SIMULATED RANDOM VARIATE FOR LANDING COST INDEX........ 92

FIGURE 4-9: SIMULATED RANDOM VARIATES FOR ENGINEERING COST INDEX........ 92

FIGURE 4-10: COMPARISON OF CONSTRUCTION COST INFLATION RATE FORECASTING 94

FIGURE 4-11: COMPARISON OF LANDING COST INFLATION RATE FORECASTING 95

FIGURE 4-12: COMPARISON OF ENGINEERING COST INFLATION RATE FORECASTING 95

FIGURE 4-13: COMPARISON OF CASH FLOW ESTIMATION - SUBJECTIVE VS. STATISTICAL (SCALED)........ 96

FIGURE A-1: MOVING AVERAGE FORECASTS AGAINST ORIGINAL DATA 110

FIGURE A-2 DATA FITTING FOR TREND REGRESSION METHOD........ 112

FIGURE A-3: FORECASTS FROM EXPONENTIAL SMOOTHING METHOD........ 114

FIGURE A-4: SCREENSHOT OF DATA INPUT WINDOW........ 115

FIGURE A-5: SCREENSHOT OF MODELING INITIALIZATION WINDOW 115

FIGURE A-6: SCREENSHOT OF ACF PLOT WINDOW 116

FIGURE A-7: SCREENSHOT OF PACF PLOT WINDOW 116
FIGURE A-8: SCREENSHOT OF MODELING PARAMETERS UPDATING WINDOW 117
FIGURE A-9: SCREENSHOT OF MODELING TERMINATION WINDOW 117
FIGURE A-10: SCREENSHOT OF FORECASTS OUTPUT WINDOW 118
FIGURE A-11: SCREENSHOT OF FORECASTS IN TIME-SERIES PLOT WINDOW 118
FIGURE D-1: INPUTS WINDOWS FOR UBJ ROOT ELEMENT 135
FIGURE D-2: INPUTS WINDOWS FOR DATA PROCESSING ELEMENT 136
FIGURE D-3: OUTPUTS WINDOWS FOR DATA PROCESSING ELEMENT 137
FIGURE D-4: OUTPUTS WINDOWS FOR PARAMETERS ESTIMATION ELEMENT 137
FIGURE D-5: OUTPUTS WINDOWS FOR DIAGNOSTIC CHECKS ELEMENT 138
FIGURE D-6: OUTPUTS WINDOWS FOR FORECASTING ELEMENT 139

LIST OF ABBREVIATIONS

ARTA: Autoregressive to Anything

ARIMA: Autoregressive Integrated Moving Average

CDF: Cumulative Distribution Function

PDF: Probability Distribution Function

SPS: Special Purpose Simulation

CEM: Construction Engineering and Management

AR: Autoregressive

MA: Moving Average

ACF: Autoregressive Correlation Function

PACF: Partial Autoregressive Correlation Function

RMSE: Root Mean Square Error

MLE: Maximum Likelihood Estimating

SSE: Sum of Squared Error

UBJ: Univariate Box Jenkins

MAPE: Mean of Absolute Percentage Error

STD: Standard Deviation

LRT: Light Rail Transit

FWM: Future Worth Method

CPI: Current Period Index

CHAPTER 1: INTRODUCTION AND LITERATURE REVIEW

1.1 Introduction

The construction industry faces the challenge of dealing with projects with substantial long durations. These projects, which include infrastructure projects such as highways, subways, large-scale storm water facilities and sewer line systems, involve a large amount of risk and uncertainty due to lengthy project duration and the variations in indoor and outdoor conditions, either physical or virtual. These factors have either positive or negative effects on project execution, and most significantly, a large number of them are time-related or time-sensitive to a certain degree. This characteristic becomes even more pronounced as the project duration grows longer.

Infrastructure project capital planning is now capturing more and more attention from the industry and the academic arena. This area of engineering application normally deals with the study of potential project requirements from medium to long range. A great deal of effort is invested in the analysis of risk and uncertainty associated with changes to project conditions along the time axis. A basic approach to predict future conditions is the use of forecasting methodologies, a group of qualitative or quantitative methods for the estimation of future outcomes. These methodologies provide deterministic forecasts for future periods at the accuracy associated with different methods in their applicable areas. However, in terms of quality analysis of future conditions for capital planning purpose, the level of accuracy obtained and the lack of risk analysis are unsatisfied and unacceptable.

The accuracy and suitability problems mentioned above can be attributed to the existence of inherent limitations in the current methods and the ignorance of more advanced methods that have better adaptability but more technical difficulties in their implementation. Therefore, a detailed application study to verify the applicability of these advanced methods and identify the solutions to overcome the implementation problems is necessary. Additionally, research must be conducted in order to develop an approach that provides the required risk analysis functionalities.

Boskers (2002) describes the use of Box Jenkins methodologies to provide more accurate and objective forecasts for future inflation rate prediction. These forecasts were then used to perform a simulation-based cash flow analysis for a transportation project with a five-year execution period. The success of Boskers (2002) has shown that the implementation of advanced forecasting methods, such as Box Jenkins methodologies, can improve the quality of forecasting results for long-term projects where the historical data is complex. However, the implementation of Box Jenkins methods described in Boskers (2002) does not solve the problem of integrating the implementation tools and the automation of the modeling process. Average users are thus prohibited from employing these methodologies in a simplified way. Additionally the risk and uncertainty associated with the forecasting were not considered for the Box Jenkins methodologies application.

Based on the previous research described above, it is necessary to perform some work to solve these problems. The related algorithms and methodologies need to be studied and the implementation tools must be designed for implementation purposes. The developed methodologies should also be verified with data from industry in the forms of validation tests and case studies.

1.2 Infrastructure System Capital Planning (ISCP)

1.2.1 Concepts

An infrastructure system comprises the whole set of public facilities at both urban and suburban levels. This set typically includes bridges, roads, highways, hospitals, schools, and subways. Transportation systems, which require large amounts of funding to construct and maintain and which last for a long time period, are the most frequently studied items in capital planning.

Capital planning and budgeting refers to the high-level study and management of capital investments over a medium or long term. Canada et al. (1996) define capital planning as "a function that takes place at the highest levels of management, such as Office of the Controller or corporate executive committee." Proper capital planning and budgeting, and management represent the top management functions of an enterprise or government and are crucial process to the maintenance of an enterprise' welfare or to a community's development and improvement.

1.2.2 Practices and Problems

Capital Planning and Budgeting involve several decision procedures. Canada, et al. (1996) list the following general procedures:

1. How the money is acquired and from what sources,
2. How individual capital project opportunities (and combination of opportunities) are identified and evaluated,
3. How minimum requirements of acceptability are set
4. How final project selections are made and
5. How post-mortem reviews are conducted.

According to the above procedures, it is certain that capital planning is performed on the condition of availability of sufficient funds. However, a number of researchers and professional groups have documented the decline of federal capital outlays for infrastructure. Woodridge et al. (2001), for example, stated that the requisite capital costs greatly exceed existing direct public funding mechanisms. Declining federal capital over the past 30 years has resulted in an increasingly constrained budget available for infrastructure revitalization. Hastak et al. (2001) emphasized that the shortage of adequate funds has imposed serious constraints on how projects are

planned, designed, constructed, operated, and maintained. Based on these comments, it is evident that decision-makers at the government level must have reliable information regarding future conditions so that the limited available funding can be directed in a reasonable manner and to the maximum degree.
It is obvious that, in view of shrinking infrastructure budget, the capital planning of infrastructure systems relies heavily on the prediction and estimation of future outcomes. Uncertainty associated with such factors as future transportation needs, future economic developments, and future market condition variations, must be estimated in sufficient detail to ensure that related planning can be founded upon a realistic and reasonable base.

Much research on capital planning systems has been conducted primarily by government organizations or research institutes. Several successful systems have strongly supported decision-making in annual infrastructure capital planning. The Infrastructure Systems Development Research Group at MIT (Massachusetts Institute of Technology), for example, is currently engaged in a long-term effort of developing computerized decision support system. "CHOICES," a decision support tool using engineering systems integration concepts, has been implemented through this project and has proven to be a successful tool (Garvin et al. 2000). In this system, estimation and forecasting are integral to the overall functionality. Historical data about government revenues, infrastructure system operating expenses are analyzed and then projected to further five to ten years. Related trend or seasonality is analyzed for further analysis. Cash flow projection for various scenarios are plotted in order to perform cost and revenue comparisons. In general, long range forecasting comprises the major part of estimating work.

There are several challenges involved in forecasting and estimation for capital planning. Canada et al. (1996) mention three: first, the unavoidability of forecasting of critical elements associated with the development of a project, second, the inherent uniqueness of each project, which makes the process of estimating unique to each project, and third, little information is available from the previous forecasting work. In response to the aforementioned challenges, however, information about past outcomes in similar and related projects may be used to estimate potential outcomes. Furthermore, the data gathered from other projects may be used to adjust project information and to assume future conditions. Many different techniques could be employed for collecting and projecting estimation data for subsequently making probabilistic estimates.

1.3 Review of the State-of-the-art Forecasting Methodologies for ISCP

1.3.1 Introduction

Successful capital planning of infrastructure systems relies heavily on the estimation of future social and economic conditions. This procedure should be an integral part of a comprehensive planning procedure (Canada et al. 1996) because it provides basic information about future outcomes. Although this procedure is very cost and time-consuming, it ultimately proves to be necessary for complete and successful planning.

Just as mentioned earlier, a big assumption in estimation and forecasting is that history will repeat itself in the future. Future outcomes must been deemed to follow certain trends that are found in the historical data as well. Much research has been organized to determine an optimal methodology for data collection and techniques for improving the quality of forecasts in order to increase the level of confidence decision-makers can have in the forecasts. Regardless of the estimation source, users should recognize that the estimate would be errorous to a certain extent; estimates and forecasts are evaluations of incomplete evidence.

Canada et al. (1996) summarized the following factors that determine the level of detail and accuracy of an estimate:

- The estimatability of that which is to be estimated
- Methods or techniques employed
- Qualifications of estimators
- Time and effort available and justified by the importance of the study
- Relevance of study results to the particular estimate

The following review summarizes the forecasting methodologies for infrastructure system capital planning, including their advantages and limitations in terms of accuracy and usability. Research representing the state-of-the-art methodologies studies will also be summarized, evaluated, and commented, in order to explain why this research is needed.

1.3.2 Current Methodologies

1.3.2.1 General Summary

Both qualitative and quantitative methodologies must be employed in order to ensure a high degree of forecast accuracy. An accurate decision must be made based upon expert judgment of the analysis results. Three technique categories have been identified as available:

1. Time-series techniques;
2. Subjective techniques; and
3. Cost engineering techniques.

The first group includes the most frequently used numerical methods in the forecasting and estimation analysis of capital planning. Among the methods available, Box Jenkins methodologies deserve particular attention because of their ability to deal with complex situations and to provide uncertainty analysis functionality. This group of methods will be explained in detail in the following paragraphs.

Subjective techniques, as the second group, represent the easiest methods of forecasting, but largely depend on the forecasters's qualification and experience , in order to generate satisfactory results. These methods, basically, utilize decision makers' personal experience and intuition to reach rough predictions about future outcomes. Therefore, these methods need to be employed together with numerical methods.

Cost Engineering techniques involve several cost estimation methods, which are used in construction industry. In general, they utilize several economic factors or indexes to predict project cost in early stage of planning considering the influence of time, geography and economic conditions. Since most conditions which influence capital system planning have a close relationship with time, it is still necessary to employ certain time-series techniques to quantity these influence, in order to come up with those factors to be used in cost engineering techniques.

1.3.2.2 Time Series Techniques

Time Series Forecasting

Forecasting attempts to predict events and conditions by examining data gathered from past projects. The intent is that by examining this information, certain descriptive patterns will emerge, informing the analyst of standard and cyclical indicative areas. The analysis of past data may be based on experience, judgements, or opinions of experts in the given field. However, the preferred means of analysis employs mathematical and statistical models. Ultimately, the primary assumption behind all forecasting techniques is that "the item to be forecast is affected by another variable, such as time, or by several variables, such as those that measure the overall economic conditions of a region" (Gaynor and Kirkpatrick. 1994).

Time-series forecasting models are based on an analysis of a chronological sequence of observations of a particular variable (Gaynor and Kirkpatrick. 1994), hence these models are called univariate models. An intrinsic feature of time-series forecasting is the dependency of adjacent observations.

Historical data plays an important role in time-series analysis, as the data collection methods will determine, to some degree, the quality of subsequent forecasts. Unreliable data can render the most sophisticated statistical techniques and forecasting models.

Gaynor and Kirkpatrick (1994) summarized various methods related to time-series forecasting as follows:

Table 1-1: Time-series Methods

Method	Description
Moving Averages	Recent values of forecasting variables are averaged to predict future outcomes
Exponential Smoothing	An estimate for the coming period based on the weighted combination of the originally forecast for the previous period and the most recent outcome.
Adaptive Filtering	A derivation of a weighted combination of actual and estimated outcomes, systematically altered to reflect data pattern changes
Time-series Extrapolation	A prediction of outcomes derived from the future extension of a least-square function fitted to a data series using time as the independent variable.
Time-series Decomposition	A prediction of expected outcomes from trend, seasonal cyclical, and random components, which are isolated from a data series.
Box Jenkins	A complex, computer-based iterative procedure that produces an autoregressive, integrated moving average model, adjusts for seasonal and trend factors, estimates appropriate weighting parameters, tests the model, and repeats the cycle as appropriate.

It should be clear that there is not an established guideline for selecting suitable time-series models for a certain practical problem. Several elements must be considered including forecasting accuracy requirement, the availability of data, the cost of collecting the data, the availability of computer programs, the time frame involved, the type of forecast desired, and, most importantly, the different mathematical assumptions each model must meet. It is obvious that a balance must be made between listed factors and the importance of the modeller's experience in making the selection should be recognized.

Time-series Simulation

Time-series simulation technologies are a functional extension of time-series forecasting. The Futures Group (1994) notes that "in simulation modeling, an attempt is made to duplicate the system being modeled in the form of equations (statistical model), not solely by drawing on statistical relationships among variables, but rather by logic and inference about how the system works." Arsham (1994) also pointed out that "System Simulation is the mimicking of the operation of a real system, such as the day-to-day operation of a bank, or the value of a stock portfolio over a time period. By advancing the simulation-run into the future, managers can quickly find out how the system might behave in the future, therefore making decisions, as they deem appropriate."

Generally, simulation techniques are divided into two main branches according to the manner in which they perform input modeling: time-independent simulation and stochastic process simulation. For review purposes, a concise illustration about time-series simulation, which represents stochastic process simulation, will be given through the discussion of their similarities and exclusive characteristics. Because time-series simulation employs similar terminologies to other simulation techniques, a basic knowledge of simulation technologies from the readers is helpful. The readers can refer to Chapter 3 for further information on specific theories, formulas, assumptions, and hypotheses, in order to fully understand the developed risk-based Box Jenkins forecasting model.

Time-series simulation collects and analyzes time-series data, which have a naturally dependent quality. Therefore, the raw data are not organized randomly but ranked in a chronological sequence. This characteristic prevents users from deleting or adding raw data points based solely on subjective rationale. Any data pre-processing should be performed with the continuity and completeness of a data set.

Different from time-independent simulation, in which input modeling attempts to fit raw data into a stochastic distribution, simulation input modeling based on time-series data considers the interdependence of data points and their sequence along a time-axis. A linear or non-linear time-series process, such as ARTA (Autoregressive to Anything) (Biller and Nelson 2002) or ARIMA (AutoRegressive Integrated Moving Average), may, therefore, provide a more suitable model for data fitting. Additionally, a standard stochastic distribution of random data, which is generated from data fitting, may combine with former time-series models to become input modeling results. These two will be input into the simulation environment finally as the random time-series process models.

Thus far, the simulation execution for both simulation techniques is similar. Traditional or enhanced Monte Carlo simulation techniques (Li and Winker 2000) are normally employed to generate the simulated output data after a certain amount of iterations. For time-series simulation, special consideration has been given to ensure that the technique is well adapted to the time-series data. In Chapter 3, readers can refer to the developed time-series simulation algorithms for these considerations.

Regarding output analysis, basic performance measures such as confidence level, probability, CDF and PDF, and percentile analysis are used by both techniques. Certain output analysis methodologies have been developed with the assumption that output data follow normal distribution. However, research studies on non-normal distributions for output analysis have been undertaken to explore more generic output analysis methods.

1.3.3 Methodology Study

This literature review indicates that several basic time-series forecasting methods, such as correlation and regression analysis, exponential smoothing, and moving average are still among the most popular tools for forecasting and estimating analysis in capital planning and continue to provide relatively satisfactory results.

Leslie and Minkarah (1997) recommended using extrapolating method to predict the yearly renewal costs of an infrastructure renewal project based on historical data; Wilmot and Cheng (2003) built a multi-regression model to predict future highway construction costs; this model was considered superior to both the unit price method and the extrapolation method for the purpose of long frame forecasting.

In studying several infrastructure capital-planning systems developed by various governmental organizations and research institutes, the author also noticed the presence of most basic time-series methods for providing time-related forecasting or estimation. Advanced techniques, such as Box Jenkins methodologies, multi-regression, and neural networks, are seldom identified or are not employed to a full degree.

Many researchers also noticed the limitations related to typical forecasting approaches. Salem et al. (2003) criticized the weakness of deterministic projection of service life of the infrastructure unit. Their research recommended a risk-based approach using probability theory and data input modeling for predicting the probability that different life-cycle costs associated with the construction rehabilitation of an infrastructure unit will occur. Wilmot and Cheng (2003) commented that the Unit Method failed to provide reliable highway construction cost forecasts due to its inability to address the variation of site conditions.

In a research performed by the Futures Group (1994), it was concluded that time-series analysis could fit historical data to a linear or non-linear function and then extend the data into the future using the function. However, this analysis cannot provide a description of how these data behave over time nor can it explicate the actual system operating beneath these data.

Hastak and Baim (2001) summarized different risk factors affecting the management and maintenance cost of urban infrastructure. They pointed out that errors in the forecasting and calculation process can lead to inadequate design criteria and may have a considerable impact upon management and maintenance costs. They recommended that risk analysis could improve the quality of information available for decision making by revealing and clarifying the implications of uncertainty through technical analysis.

A number of researchers have studied the forecasting features of Box Jenkins methodologies. Several successful application studies have been performed proving Box Jenkins methods to be a flexible and powerful tool for short and long range forecasting tasks. It is necessary to emphasize that these studies on Box Jenkins methodologies are typically found in economics, financial, and social sciences areas; these areas are more academic and advanced in application. In the area of engineering, a literature review identified a few studies on the use of Box Jenkins methodologies. These studies concern primarily hydraulic engineering and ecological engineering. Sun and Koch (2001), for example, built a Box Jenkins ARIMA model to analyze and forecast salinity in Apalachicola Bay, Florida. Li et al. (2003) employed Box Jenkins models to predict an index of soil dryness in the southwest of Western Australia.

Boskers (2002) applied Box Jenkins forecasting methodologies in his study of uncertainty in capital construction project planning in order to predict future inflation rates so that related cash flow analyses can be performed accurately with more trustworthy inputs on future estimation than those which were based solely on assumptions from experience.

In spite of these advancements, risk analysis associated with forecasting and estimation has not been identified and explored to its full extent by the academic or industrial communities with regards to capital planning for infrastructure system. However, several researchers in other areas of engineering have employed general time-series simulation concepts to analyze risks associated with practical problems.

Harding and Walski (1996) combine time-series analysis and simulation analysis to study contamination in the water distribution system of a major city. They found that time-series analysis provided a continuous representation of the conditions but, without the application of simulation, the time-series analysis cannot lend itself effectively to statistical analysis, especially when these analyses extend to a long time. Bashford et al. (2003) tested an Even Flow production methodology for the U.S. Housing Industry with the assistance of time-series simulation. A *Simphony* SPS template was developed for the purpose of implementing purpose. Enright and Frangopol (1998) performed a Monte-Carlo simulation based on time-series data to find the probability of a cumulative-time system failure for deteriorating concrete bridges.

1.3.4 Limitations of Current Methodologies

A literature review of infrastructure capital planning practices and current methodologies for the quantitative forecasting and estimation of infrastructure capital planning indicated several limitations, which have seriously hindered scientific forecasts and uncertainty analysis from becoming a basis for successful capital planning. These limitations have already affected the overall management of infrastructure systems in an adverse manner.

1.3.4.1 Theoretical Limitations

Current methodologies for forecasting and estimating in capital planning emphasize the simplification of algorithms, while are incapable of providing accurate and meaningful forecasts with associated risk analysis. This problem is partially attributed to the limitations of current methods in adaptability and accuracy. Another reason is that current methods are incapable of producing risk analysis information due to the limitation of their mathematical and statistical algorithms. From the literature review, some limitations associated with current methods are summarized as follows:

Regressive analysis is deemed to be a powerful tool for predicting future outcomes based on independent influencing factors. However, regressive analysis requires a pre-defined function format as the default relationship model and uses sample data to verify it. Thus this approach proves to be a subjective method. Similar research endeavor has failed to find a suitable function because the parameters related to the function are either difficult or impossible to estimate. Although sample data can be used to verify a function, the results may still fail to represent a general condition. Any outliers existing among the sample data can nullify the whole estimate. The model that results from a regressive analysis often loses its usability when compared with objective judgments. Another weakness of the regression model is that it assumes statistically independent data, whereas time-series data often possess an interrelated quality.

Exponential smoothing is superior to other basic methods in its accuracy. It is also more sensitive to future change than linear regression (Canada et al. 1996). However, it uses data inefficiently for initial modeling. It also assumes that data to be used are statistically independent. For dealing with actually dependent, time related data, this method is not appropriate.

Sample moving average method provides the worst accuracy (Armstrong 1985). It also uses the data inefficiently for initial modeling. Similar to exponential smoothing in basic algorithms, it also only applies to only independent data, which are seldom identified in most capital planning work.

Regarding the mathematical and statistical algorithms of current methods, the author noticed that current methods have not provided well-developed risk analysis function to support forecasting. The author believes that this problem results from the difficulty in functionality extension due to the low level algorithms associated with current methods and the lack of related research endeavors.

1.3.4.2 Practical Limitations

This literature review found that current methodologies, although frequently used for forecasting and estimation tasks, perform poorly overall as they did not provide sufficient information to meet the requirements of current complex and comprehensive decision supporting. Deterministic point estimates or simple range estimates constitute the extent of the information that these basic methods can provide. More statistics and uncertainty analyses are required from the industry. To implement this, current methods should be refined or new advanced methods should be explored. However, a practical limitation associated with some advanced forecasting methods, such as Box Jenkins, is that these methods are seldom used due to the complexity involved in implementation.

One of the reasons for their limited use is that the application of advanced methods requires more mathematical and statistical knowledge than the basic methods. Average engineers, however, may not possess this knowledge. Another reason is that advanced methods normally need more time and effort than basic methods. Often, specialized modeling experience is also needed.

Both the industry and the academy have recognized the importance of time-series simulation in analyzing uncertainty problems associated with forecasting tasks. However, until now, no mature simulation modeling methodologies had been developed that could provide engineers with a generic approach for simulation analysis based on time-series data. Although other modeling methodologies in areas such as finance, economics, social sciences, have been successfully explored and implemented, neither their time-series process models, nor their modeling methodologies have gotten proved to be applicable in engineering.

1.4 Research Objectives and Methodology

In order to not only solve the practical problems of advanced forecasting methodologies, but also contribute to the academic state-of-the-art in the field, the following research objectives will be followed:

- A detailed application study on Box Jenkins methodologies will be performed and an integrated, robust, and automated Box Jenkins modeling tool will be developed in the *Simphony* simulation-modeling environment as a SPS (Special Purpose Simulation) template.
- A whole set of input modeling methodologies for univariate random process modeling, represented by univariate time series process simulation modeling, will be developed along with associated implementation tools.
- The incorporation of risk analysis functionalities into forecasting methodologies will be studied. Risk-based forecasting analysis will also be implemented in the developed methodology.

In order to achieve these objectives, this research work is broken down into the following steps:

- Advanced forecasting methodologies, specifically Box Jenkins methods, will be studied in detail with a focus on its application to capital planning of infrastructure system. It is also essential to automate and integrate the modeling process.

- The risk and uncertainty will be studied and overcome by incorporating risk analysis functionalities in the proposed forecasting model.
- Univariate random process modeling, specifically univariate time series will be studied and developed to support risk analysis for forecasting methodologies.

1.5 Conclusion

The execution of civil projects with a lengthy duration is problematic for the construction industry because the risk and uncertainty inherent to these projects is normally larger than other projects. To solve this problem, construction management professionals can employ forecasting methodologies to better understand future outcomes. However, the methods, that are currently available, do not provide a satisfactory level of accuracy for complex project environments and attempts to implement more adaptable and powerful methods have proven to be very difficult. More importantly, expected risk and uncertainty analysis is seldom identified or even carried out. In short, the requirements from the industry cannot be satisfied using the existing forecasting methods.

This chapter reviewed the concepts and practices of the current level of infrastructure system capital planning and related forecasting methodologies. It also gave special attention to one group of methodologies, that is, time-series techniques, through a brief introduction to time-series forecasting and time-series simulation. A review of the state-of-the-art in research on forecasting methodologies in infrastructure system capital planning helps to identify the limitations that exist at both academic and practical levels. This review demonstrates expressly the importance of the accurate estimation about future outcomes for successful capital planning. The insufficient application of advanced forecasting methodologies, however, may also inhibit the delivery of satisfactory estimation data and disable potential future planning and budgeting.

This research, therefore, aims to solve these problems and provide usable implementation tools to the industry. It also has the objective of providing a theoretical framework to solve the problem using input modeling algorithms for univariate time series process for simulation modeling.

CHAPTER 2: AUTOMATED BOX JENKINS FORECASTING MODELING

2.1 Review of Box Jenkins Methodologies

Before a detailed introduction may be undertaken, some basic Box Jenkins concepts and theories must be reviewed. The explanation of a series of standard notations will help to facilitate the comprehension of the following illustrations.

Y: original time-series data;
T: time, from one to n;
ε_t: random error in time period t, t=1 to n (also called white-noise forecast error);
Z: stationary time-series data;
μ: the constant coefficients ARIMA model function;
Θ_i: moving average coefficients, $\iota = 1$ to q;
ϕ_j: autoregressive coefficients, j=1 to p;
p: order of autoregressive model;
q: order of moving average model; and
δ: constant value

In theory, Box Jenkins models can only process stationary time-series data. A stationary time-series is a series of univariate data that does not contain trends or seasonality, that is, it fluctuates around a constant mean. However, the original data collected from real jobsites are virtually non-stationary. Therefore, certain transformation methods should be applied in order to transform original data into stationary data. Frequently used transformation methods include n-order normal differencing, n representing the order of differencing, n-order seasonal differencing, n representing the order of seasonal differencing. Natural logarithms transformation is another transformation to reduce the variation in variance. There have been other transformation methods besides these three, however, which are seldom used in engineering problems.

After an iterative transformation process on the original data, one can obtain statistically stationary time-series data. Subsequently, in most cases, a fitted ARIMA model should be identified to represent the linear or non-linear relationship within the current situation and among past situations, including past forecasting errors. The following function shows the ARIMA model:

$$Z_t = \delta + \phi_1 Z_{t-1} + \phi_2 Z_{t-2} + \phi_3 Z_{t-3} + \cdots + \varepsilon_t - \Theta_1 \varepsilon_{t-1} - \Theta_2 \varepsilon_{t-2} - \cdots \text{(Equation 2-1)}$$

Where:

Z_t = the present stationary observation; and

$Z_{t-1}, Z_{t-2}, \ldots, \varepsilon_{t-1}, \varepsilon_{t-2}, \ldots$ = the past observations and forecasting error for the stationary time series (not original time series); and

ε_t = the present forecasting error (the expected value is set equal to 0).

Depending on the complexity of the real system under analysis, the above ARIMA function may be formulated as simple or complicated, which means that less or more past observations or forecast errors will be included in order to calculate the current condition.

After identifying the above function for a specific problem and estimating all potential model parameters, one will arrive at a deterministic relationship function. Moving the point in time one unit ahead, a relationship is obtained among future conditions and current available information. Thus, the fitted ARIMA model function can be used to forecast future outcomes.

It should be noted that an estimated future value comes from a calculation using stationary time-series data. Therefore, this is an estimate not a real future value. To retrieve the original value, which concerns a series of back-transformations must be performed. These operations mirror the operations of former transformation endeavours.

In the ARIMA model function, ε_t, the present forecasting error will be transformed to ε_{t+1} when the time point is moved one unit ahead. In this situation, it is assumed that the error is approximately equal to zero when one wants to get point estimate value. Therefore, for the calculation of future value, one must replace ε_{t+1} with zero.

All above content comprises the basic theories and methodologies of Box Jenkins Forecasting. To understand the research involved in automating Box Jenkins forecasting modeling, and the following proposed risk-based Box Jenkins forecasting model discussed in Chapter 3, interested readers can refer to literature on Box Jenkins methodologies. Both *Forecasting with Univariate Box Jenkins Models-Concepts and Cases* (Pankratz 1983) and T*ime Series Analysis - Forecasting and Control* (Box et al. 1994) provide a helpful general review about time-series techniques, which include Box Jenkins methods.

2.2 Rationale for Selecting Box Jenkins Methodologies

In order to fully comprehend a discussion of the research, it is first necessary to explain why Box Jenkins methodologies, in particular, were chosen for detailed study and the overall rationale behind this selection.

Based on an analysis of the limitations found in current methodologies, as discussed in Chapter 1, the author determined the necessity for exploring more advanced methods of forecasting and estimation. The criteria, by which the author ultimately selected the Box Jenkins methodologies, included:

- The capability of dealing with complex situations;
- The adaptability of processing dependent time-series data;
- The level of advancement in mathematical and statistical backgrounds;
- The existence of a risk and uncertainty analysis functionality; and
- The simplicity of its implementation.

Box Jenkins methodologies often produce, in a great many cases, the most accurate forecasting models for any set of data (Gaynor and Kirkpatrick 1994). These methods also provide a more systematic approach to building, analyzing, and forecasting time-series models. Armstrong's comparison test (Armstrong 1985) on the ranking of extrapolation methods (from the highest rank as 1 to the lowest rank as 5) in terms of cost, understandability, and forecast accuracy for both short range and long range ranked Box Jenkins methods as 1.5 for short-range forecasting accuracy and 2 for long-range forecasting accuracy.

Box Jenkins models are capable of dealing with those dependent time-series data. These data, in theory, are not suitable for other methods. For example, a regression model has a standard assumption that the error term should be statistically independent. In reality, much time-related data are dependent or have an autocorrelation among them.

Generally, Box Jenkins methodologies use the most recent observations as starting values and then analyze recent forecasting errors to determine the proper adjustments for future periods. In doing so, they allow timely adjustments of error levels and provide a more flexible imitation of a particular complex trend or seasonality.

From the following research discussed in Chapter 3, it is observed that Box Jenkins methodologies can be extended to incorporate risk analysis functionality into traditional forecasting functionality using time-series simulation technology. This feature enables the Box Jenkins methodologies to overcome the problems that exist with deterministic point forecasts and simple range forecasts, by providing more statistical results, such as percentile analysis and probability estimation. Thus the uncertainty associated with point forecasts and simple range forecasts can be analyzed in a more systematic and comprehensive way.

Questions may arise regarding the methodology's usability and simplicity for use in solving engineering problems. These questions arise out of a belief that a successful forecasting support model should be advanced and theoretically founded, yet practically easy-to-use for average users. Box Jenkins methodologies prove to be among the most advanced and useful univariate forecasting models. This methodology can also be simplified according to the specific requirements of a certain application area such as capital planning and can, therefore, be implemented by a specially designed and developed computer program, as discussed in the following sections. The application of these complicated methodologies, therefore, can be easy for engineering professionals. The author's research about the automation of the Box Jenkins modeling process using computer simulation technologies has solved this problem.

2.3 Implementation Problems Associated with Box Jenkins Forecasting Modeling

Thus far, previous research has shown that Box Jenkins methodologies are superior to other forecasting methods because of their comparably excellent estimation accuracy for short and medium range (one to three years) forecasting and their exclusive functionality in risk and uncertainty analysis, which becomes increasingly important for complex forecasting and planning work. However, according to the author's literature review, there are several significant problems in using this method, which still prohibit extensive application of the method in the engineering domain.

1. Box Jenkins methodologies are more complicated and involve several iterative procedures, which require cumbersome calculations, compared with other methods. There is no automated procedure to facilitate the modeling procedures.
2. Box Jenkins methodologies comprise a family of Autoregression and Moving Average Integration (ARIMA) models, from which the user chooses the model most suitable to their projects. They allow for a high degree of flexibility in the selection of a model. However, they also call for much subjectivity on the part of the analyst. This requirement means more experience and subjective decision-making will be needed.
3. It is difficult to update a fitted Box Jenkins model. As new data are acquired, the model must be rebuilt, which means a sharp increase in the amount of modeling work.

Several researchers have commented upon these problems specifically in terms of their concerns regarding the application of this method. In Armstrong's study on extrapolation methods (1985), he summarized that Box Jenkins procedures are still more difficult to understand than other methods because of their complexity even with the rapid development of computer technologies. Tang et al. (2003) compared Box Jenkins models with other methods used in forecasting Annual Average Daily Traffic (AADT) in Hong Kong. They also mentioned that the ARIMA modeling requires extensive data calibration.

2.4 Automating Box Jenkins Forecasting Modeling

With the advance of computer technologies, much of the reasoning and logical thinking, originally undertaken by humans, can be replaced by computer programs. Tedious and iterative data manipulations can be executed at an astonishing speed by computers. It is obvious, therefore, that through a careful study of the methodologies and creative programming, the Box Jenkins forecasting modeling process, which is often quite challenging to average users, can be performed automatically with little to no difficulty using a computer program specially designed for that purpose in particular.

In general, the automation and integration of Box Jenkins modeling processes are achieved by a successful simplification of classical Box Jenkins methodologies for specific engineering areas and with the aid of computer technology. This approach arose came partly from a thorough study of Box Jenkins methodologies, including classical concepts and theories, modeling procedures, currently available commercial software, and other similar works, and partly from the analysis of engineering-related problems, including issues dealing with particular engineering characteristics, data types, case studies, and research reviews.

It should be stated that a universal and generic solution is, thus far, an impossible task in theory: Box Jenkins forecasting methodologies are impossible to be automatically executed in full. Difficulties exist within the theories themselves and as a result of the necessity for human input during the iterative process. It is obvious that human experience and judgments cannot be simulated completely, which precludes an absolutely automated process without any human involvements.

In order to deal with the problems involved in implementing Box Jenkins forecasting and to provide a systematic and automatic modeling tool that does not require too much human involvement or experience, a summer project was initiated at the beginning of May 2003 in Construction Engineering and Management (CEM) group of University of Alberta to perform research on the automation of Box Jenkins modeling processes. This project aimed at exploring solutions for overcoming the weaknesses mentioned above; to study the modeling procedures involved in discrete-event simulation; and finally to develop a Special Purpose Simulation (SPS) template to simulate and automate the modeling process in *Simphony*, a simulation environment.

Much research and many case studies have been reviewed in an effort to put together the experiences and pitfalls of the modeling process. The statistical and mathematical theories of these methodologies have been checked to determine ways of simplifying modeling procedures. A pilot SPS template was developed using *Simphony*. Several artificial case studies were conducted to verify the functionality of this template and the results of these studies prove that this template is capable of performing, to a satisfactory degree, automatic Box Jenkins modeling process.

The foundational ideas of this template are as followings and the author assumes readers' knowledge of *Simphony* and SPS development terminologies and techniques.

1. Original data processing operations, such as normal differencing, seasonal differencing, and natural logarithm transformation, can be performed by computer programs. All related calculations and graphical representations can be performed using *Simphony* SPS template with interrelated functional elements.

2. One of the most difficult and time-consuming modeling procedures, the identification of the stationarity of time-series data can be simulated by developed computer codes using an iterative judgment mechanism.

   The literature review shows that for average time-series data, normal differencing with 3 or less order, seasonal differencing in 4, 6, 12 and 24 lags and natural logarithm transformation, are mostly capable of processing the raw data to stationary data with satisfied performance. In this template, all possible combinations of the processing methods mentioned above will be attempted. T test statistics for Autoregressive Correlation Function (ACF) and Partial Autoregressive Correlation Function (PACF) for a processed time series using each combination of methods will be calculated to compare with predetermined critical values. Theoretically, the results of comparison between T test statistics and predetermined critical values will determine the point of termination for the raw time-series data processing procedure.

3. A suitable ARIMA model identification process can be simplified through enumeration methods.

   Many articles written about Box Jenkins modeling summarize that the orders of Autoregressive model and Moving Average model, named p and q are usually less than or equal to two (Gaynor and Kirkpatrick 1994). The author, therefore, make the assumption that, for most cases, when Box Jenkins methodologies are used, fitted ARIMA models do not have more than three order of AutoRegressive model (AR) and three order of Moving Average model (MA). With this assumption in mind, it is easy to simplify the process by enumerating all possible combinations of ARIMA models with p and q as less than 3. The total number of possible combinations – ARIMA (p, q), p, q =0 to 3 but p and q cannot be zero at the same time – is calculated as 15. After estimating the parameters for each model and performing the necessary diagnostic checks on them, one should be able to determine the most suitable model, which has, in its simplest form, the least Root Mean Square Error (RMSE) and passes all diagnostic checks, which will be introduced in the following sections.

4. After the ARIMA model has been identified, forecasting, including the required reverse processing of forecasts values, can be performed by developed computer codes. Those codes can be designed as interrelated functional elements within the *Simphony* SPS template.

These ideas can be implemented in an integrated way by developing the necessary algorithms within a *Simphony* SPS template. This template contains all the functional elements for the different available modeling procedures. The expected integrated and automatic Box Jenkins forecasting modeling can then be realized.

2.5 Implementation of Automated Box Jenkins Forecasting Modeling

A contribution of this research to the state of the art in time-series forecasting techniques was the development of the integrated automatic Box Jenkins forecasting modeling system. Using this system, average users can employ the method without the hassle of comprehending complex mathematical and statistical knowledge or the requirement of extensive modeling experience. The most valuable gain emerging from this research is that the total time and cost spent on the complete modeling process and in subsequent forecasting work can now be dramatically decreased.

In general, this system employs the powerful computing capabilities of *MATLAB*-based computer programs in order to realize the algorithms associated with Box Jenkins methodologies. These programs are integrated into a computer-modeling tool, which is represented by a SPS simulation template in *Simphony*. All the elements in this template can be used together or independently to accommodate specific modeling requirements. In this way, the traditional manual modeling can be simulated by a highly efficient computer simulation program resulting in better performance with less cost and in a shorter time period.

2.5.1 Core Algorithms Design and Programs Development

The required algorithms associated a complete execution of Box Jenkins forecasting modeling are identified and summarized as following:

- Natural logarithm transformation
- Normal differencing
- Seasonal differencing

- ACF calculation
- PACF calculation
- T test
- Q test
- Parameters estimation
- Forecasting
- Inverse normal differencing
- Inverse seasonal differencing
- Inverse natural logarithm transformation

Detailed explanations regarding the above algorithms can be found in the available literature on Box Jenkins forecasting methodologies.

To realize the computerized modeling process, one must determine a way, first, to execute these algorithms in a computer environment. Second, the algorithms must be implemented with computer codes and an interface for communication between algorithms must be designed. In this way, the algorithms can work together to generate the expected results.

The above algorithms were coded using *MATLAB* programming language. The selection of *MATLAB* as the development platform was due mainly to the experience *MATLAB* has in packaging many mature mathematical and statistical algorithms in different application areas.
All *MATLAB* programming codes for those algorithms can be referred to in Appendix C, which also provides simple explanations of the codes. In order to understand these codes, certain details are necessary regarding *MATLAB* programming, the algorithms themselves, including the input and output, and Box Jenkins methodologies. Complete explanations, however, on these topics are beyond the scope of this thesis.

2.5.2 Program Integration – Development of Automated Modeling Tool

The computerized execution of Box Jenkins forecasting algorithms can be realized through the research discussed in the previous section. The traditional manual modeling process can be assisted by these programs. In this process, however, there is still a lot of time spent on the iterative selection of the best ARIMA model and upon an iterative judgement of time-series stationarity, during which extensive human experience and judgements are employed to facilitate the execution.

The following Figure 2-1 illustrates the traditional modeling process. Based on the analysis of this process, two modeling procedures – time-series stationary transformation and best ARIMA model identification - prove to occupy the most portion of modeling time.

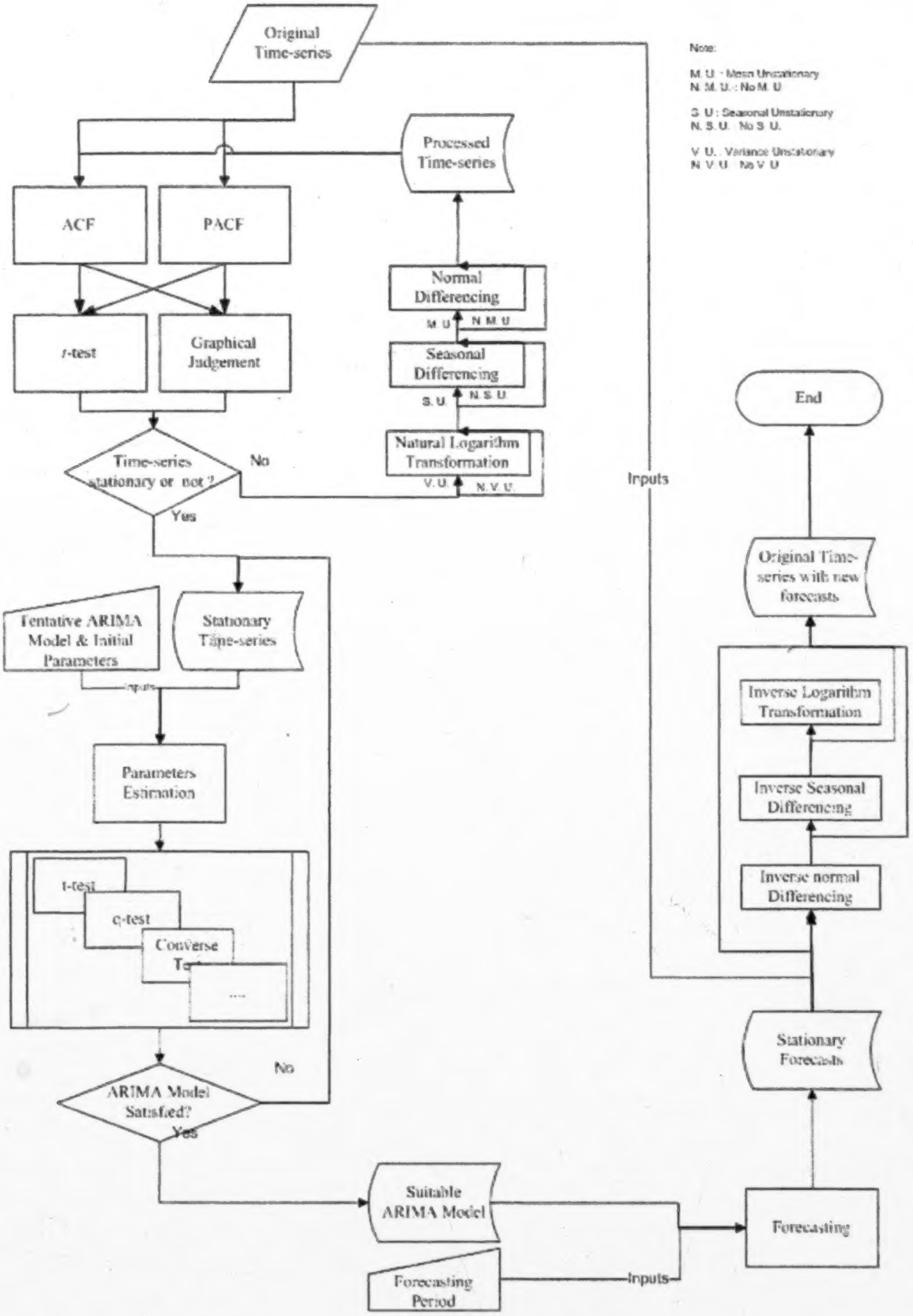

Original Time-series
Processed Time-series
ACF
PACF
t-test
Graphical Judgement
Time-series stationary or not ?
No
Yes
Normal Differencing
M. U.
N. M. U.
Seasonal Differencing
S. U.
N. S. U.
Natural Logarithm Transformation
V. U.
N. V. U.
Note:
M. U. : Mean Unstationary
N. M. U. : No M. U.
S. U. : Seasonal Unstationary
N. S. U. : No S. U.
V. U. : Variance Unstationary
N. V. U. : No V. U.
Inputs
End
Original Time-series with new forecasts
Inverse Logarithm Transformation
Inverse Seasonal Differencing
Inverse normal Differencing
Tentative ARIMA Model & Initial Parameters
Stationary Time-series
Inputs
Parameters Estimation
t-test
q-test
Converse Test
....
ARIMA Model Satisfied?
No
Yes
Stationary Forecasts
Suitable ARIMA Model
Forecasting
Forecasting Period
Inputs

Figure 2-1: Flow Chart of Box Jenkins Forecasting Modeling Algorithms

It is obvious from an analysis of this diagram that there exists the possibility of integrating all these algorithms in order to realize a more automatic and timesaving modeling procedure if a special integration system is developed. Based on the discussion of foundational ideas of a proposed SPS template, the following simplifications and modifications to rational modeling methodologies are developed to facilitate the development.

I. Time-series Stationary Judgement

Based on the foundational ideas discussed in Section 2.4, developed MATLAB programs for data transformation and for hypothesis test are utilized to perform computerized stationary judgement. Instead of applying all available transformation methods, this template will use a specifically selected set of methods with limited method parameters to perform data transformation process. After each transformation, a certain hypothesis test will be applied to test the stationarity of time-series data, in order to decide whether to terminate the judgement process or not. All these procedures can be integrated by proper computer programming.

II. Best ARIMA Model Identification

Best ARIMA Model Identification, can be described as the identification of a fitted ARIMA model and the estimation of model parameters. This process, therefore, involve the application of many mathematical and statistical methods, much modeling experience from modellers, and highly involvement of human judgement. However, from the perspective of automated forecasting modeling, traditional methods should be modified to facilitate the implementation of this procedure in a timesaving and robust approach.

Traditional methods emphasize the identification of an individual model. A suitable ARIMA model is identified through a time consuming and iterative cycle of selection, estimation, judgement and reselection. This approach causes time inefficiency and trouble for the implementation in the computer program.

Biller and Nelson (2002) undertook research on the development of input modeling algorithms for the fitting of time-series input processes for a simulation experiment. They experienced a similar problem associated with the implementation of fitting an individual ARTA (Autoregressive To Anything), which is a time-series model similar to ARIMA model, onto a Johnson translation system. If the traditional approach is taken, then all family members of the Johnson translation system should be fitted and the associated goodness of the fits should be performed by each trial. They observed that this approach works only when the autocorrelation among samples is relatively weak, and when the size of the samples is large enough. However, when the strength of the autocorrelation increases, they observed a significant bias and variability in the higher sample moments, increasing the likelihood of identifying the wrong model. Thus, they adopted an approach that would fit all of the models and compare the goodness of fits.

This creative approach inspired and verified the author's thinking in the selection of a suitable ARIMA model. Similarly, the author operates based on the assumption that ARIMA models with an autoregressive order less than three and a moving average order less than three are able to represent nearly every real system found in engineering problems. This assumption was verified by an analysis of average construction management problems and suggestions from experienced forecasters and from simulators. Gaynor and Kirkpatrick (1994) also supported this assumption by demonstrating that in most cases the autoregressive order and moving average order are not more than two.

Considering all possible combinations of these two sub-models, the total number of ARIMA models is 15. Our approach, in short, fits all 15 ARIMA models to stationary time-series data, then estimates the parameters and employs a fitting procedure, described as follows, to determine the most suitable model. This approach, although it involves more calculations than the traditional individual model approach, creates an effective solution for computer implementation. Furthermore, the increases in calculations are addressed immediately using a computer familiar in dealing with repetitive and iterative calculations. This finding, therefore, greatly facilitates the developing of automatic modeling algorithms.

The detailed mathematical and statistical algorithms for the estimation of model parameters are quite complex, involving the employment of several tools, which include Maximum Likelihood Estimating (MLE), non-linear optimization, matrix - based calculation, and matrix manipulation. A detailed development of algorithms was performed using MATLAB 13.0. This program was chosen as the development platform due to its history in designing and packaging algorithms using powerful and customizable toolboxes.

A constrained non-linear optimization is performed in order to estimate the ARIMA model parameters. This algorithm needs a vector of initial parameter estimates in order to start the iterative calculations; otherwise, convergence problems may arise. Gaynor et al., (1994) recommended the following start values for ARIMA model parameters:

For ϕ_1, ϕ_2, Θ_1, Θ_2: use 0.1 as the preliminary estimate;

μ: an estimate of μ is $\hat{\mu} = \overline{Z}$, where $\overline{Z}$ is the mean of the stationary time series;

δ: an estimate of δ is $\hat{\delta} = \hat{\mu}\,(1-\hat{\phi}_1 - \hat{\phi}_2 - \ldots)$, where $\hat{\phi}_1, \hat{\phi}_2 \ldots = 0.1$.

There are two conditions related to the selection and calculation of parameters.

Stationarity Condition: the sum of the coefficients or of the parameters of the AR model should be less than one.

Invertibility Condition: the sum of the coefficients or of the parameters of the MA model should be less than 1.

These two conditions apply for both the initial parameter estimates as well as for the final estimates. An explanation of the initial parameter values is easier to understand with this in mind,

A detailed explanation regarding fitting algorithms, however, is beyond the scope of this thesis. Further information about each algorithm can be found in the references literature. In the following chapter, the developments of MATLAB programs for all the algorithms are discussed and complete MATLAB Syntaxes for the performing of ARIMA model fitting are listed in an appendix. The programs associated with the data fitting are packaged into the developed SPS *Simphony* template and represented by a template element. This element can perform automatic data fitting, which reduces the overall workload. The related development details and computer codes are also provided in the following chapter and in the appendix.

III. Diagnostic Checking

Diagnostic checking is a complicated procedure in Box Jenkins forecasting modeling. Some conditions applied in this procedure are also employed to test the stationarity of time-series data. Therefore, to successfully modify or simplify this procedure will contribute a lot to the development of an automated modeling tool.

Normally, one performs the following procedures to check the fitted ARIMA models:

- **The iterative process must converge.**

  This requirement refers to the minimization of Sum of Squared Error (SSE) during the estimation step. If there are no other parameters (with a relative change of less than 0.001) yielding a smaller SSE, then the data fitting process is deemed to converge. This condition checking can be implemented using the developed *MATLAB*-based computer program – Parameters Estimation – discussed in Section 2.5.1

- **The invertibility and stationarity conditions should be satisfied.**

The above computer program can implement the checking of these two conditions too.

- **The residuals or forecasting errors should be random and approximately of normal distribution.**

For this condition, one either analyzes the residuals by plotting the ACF and PACF of residual series to observe whether these functions are statistically equal to zero or not, or perform certain hypothesis tests to assist in this diagnostic check. Modified Box-Pierce statistics, for example, can be used for this purpose.

To facilitate an automated checking process and reduce as much as possible human involvement, this checking is implemented by the developed MATLAB program – Q test, which is discussed in Section 2.5.1. In addition, the following simplification is made:

Usually this statistic is computed for lags 12, 24, 36, and 48. In order to simplify, however, the check on Q statistics for lag k=12 or k=24 is sufficient.

- **All estimates of parameters should be significantly different from zero (significant t-ratios).**

Using the same concept in regression analysis, the significance of each parameter must be checked. The test of significance is the *t*-test:

$$t = \frac{\text{point estimate of parameter}}{\text{standard error of estimate}} \text{ -------------------------------- (Equation 2-2)}$$

If the t ratio for a parameter is significantly greater than a predetermined critical value (usually $|t|>2$, for $\alpha=0.05$), it is retained. Otherwise, the ratio ought not to be employed and the model requires a recalculation using the remaining terms.

This condition checking is implemented by a developed *MATLAB* program – Parameters Estimation.

- **The model should be presented in its simplest form**

Past research and application studies have proven that the best Box Jenkins model always has the least number of parameters. This paradigm arises not only due to the ease in fitting and explanation of simple models, but also since parsimonious models avoid the problem of parameter redundancy.

This condition applies when a traditional individual model identification approach is used to select ARIMA models. Since in this case, only 15 low order ARIMA models are chosen for real data fitting and for the parameters estimating, and since these models are already in simplified form, a diagnostic check of this condition is, therefore, not needed.

- **The model should have a small Root Mean Square Error (RMSE)**

Several good models can be generated even after all of the above checks are completed. To identify the best of them, one should check the RMSE. The model with the smallest RMSE is the best model.

$$SSE=\sum_{t=1}^{n}e_t^2 \text{ -- (Equation 2-3)}$$

$$\text{RMSE}=\sqrt{\frac{SSE}{n-n_p}}=\sqrt{MSE} \text{ -- (Equation 2-4)}$$

Where:

n=the number of data points in the stationary time-series;

n_p=the number of parameter estimates in the model; and

e_t =the forecast error in time period t.

This condition regarding the significance of the smallest RMSE is used in our approach as the final criteria for terminating the iterative model identification process. This consideration is very important for designing automatic computer modeling. The selected ARIMA model, which will be among the 15 tentative models and which passes all the diagnostic checks, should have the smallest RMSE.

Based on the above modifications and simplifications, a more robust and easy-to-use computer system, operating with simulation technologies in order to mimic the original time-consuming manual modeling, was developed in *Simphony*, in an effort to explore this possibility. The central objective of this development is to integrate all the algorithms into one SPS template with interrelated functional elements, so that the modeling process can be performed automatically using a pre-designed simulation experiment. These functional elements are represented by virtual simulation elements, within the *Simphony* simulation environment.

The following paragraphs introduce the conceptual development of this SPS template.

Template Structure Design

The *Simphony* SPS template was developed for special purposes in simulation experimentation. To facilitate the use of the SPS template in a simulation environment, the template will usually adopt a basic father/child hierarchical structure, and package all the functional elements under one father element. In this way, the management of the developed simulation model is simplified and the structure is given an aesthetic order.

The developed SPS template for integrated and automatic Box Jenkins forecasting modeling has a father element, "UBJ (Univariate Box Jenkins) Forecasting Root", which serves as a gate for original time-series inputs to enter and from which Box Jenkins forecasts may emerge. The simulator can input historical data and specify forecasting requirements, such as the number of periods to forecasts. The forecasts originate in the child model, which is constituted by child elements, thus the father element also serves as the medium for the display and storage of forecasts from the child model.

The child elements have four members in total. They serve different functions, which in Box Jenkins forecasting terms, consist of different modeling procedures. During the modeling process, these four elements communicate with each other and transfer data and information back and forth in order to realize the iterative process.

The following table demonstrates the corresponding relationship between element and functionality:

Table 2-1: Child Elements and Corresponding Functionality

Element	**Functionality**
UBJ Forecasting Root Element	Historical Data Input and Forecasts Display and Storage
Data Processing Element	Data Transformation
Parameters Estimation Element	Models Parameters Estimation
Diagnostic Check Element	Selection of Best Model
Forecasting Element	Forecasts Generation

Algorithms' Design and Reference

As mentioned above, each of the five elements has a specific functionality. All the elements can be used together to constitute a Box Jenkins Forecasting Model for performing an automatic modeling process. The most important part of this development, therefore, is to design a template for the realization of these functionalities and to facilitate the communication and cooperation among these elements.

In *Simphony* SPS template development, *Simphony* uses Sax Basic as its programming language. This language has the similar syntax to Visual Basic, another advanced programming language; thus in theory, any functionality can be realized through proper Visual Basic programming.

In order to realize the functionalities for automatic Box Jenkins forecasting modeling, the programmer must design programs for each of the algorithms summarized in the previous sections. The computerized design of algorithms proves to be difficult even for the professional mathematician. In this research, therefore, the author adopted a straightforward process in which, that is, the algorithms were designed using *MATLAB* programming language and integrated these algorithms into a *Simphony* SPS template. The most difficult stage of the programming can be performed by a third party tool. The remainder of the work involves seeking a means of referencing these algorithm programs in *Simphony* and how to facilitate the communication among them.

The solution for these problems is to package and compile all the *MATLAB* algorithms into a third party Type Library file and design the *Simphony* codes to refer to these library files for the realization of these algorithms. The implementation of this solution is not as difficult as one initially supposed, due to the powerful toolbox offered in *MATLAB* 13.0, which has the exact functionality one required for the packaging and compilation of algorithms programs. *MATLAB* COM Builder, which is a special toolbox used for the creation of stand-alone computing programs, was provided with *MATLAB* 13 Version 0. This toolbox uses a simple operation interface to facilitate complicated COM file compiling process, in order to enable the non-professional programmer to create COM files in a professional way.

In the previous section, the author indicated the choice to design all the algorithms in *MATLAB* programming language for Box Jenkins forecasting modeling. As a result, the *MATLAB* COM Builder toolbox may be used to package these algorithms files into one deployable COM file. All the algorithms, after compiling, will become built-in functions to which *Simphony* development codes will refer.

The detailed instructions for the usage of this toolbox and for referencing the compiled COM files for the Visual Basic programming environment are provided in the *MATLAB* 13.0 Products Help files. Example codes can be found in Appendix C.

Template Development and Programming

The SPS template for automatic Box Jenkins forecasting modeling was developed with the help of the "*Simphony Developer' Guide*" and "*Simphony User's Guide*". All programming codes comply with Sax Basic programming standard, which are similar to those in Visual Basic programming. The complete program codes, however, are not attached because of the page limits requirements. Interested readers refer to the SPS template for detailed coding and explanations. Figure 2-1, illustrated in the previous section, demonstrates the general data flow and related algorithms.

In order to facilitate communication and cooperation between different elements in this template, certain public variables are set up for storaging the interim results from each element. These results occur when a certain kind of operation or calculation is performed. They are retrieved from these public variables when needed by a successor element for its modeling procedure.

Appendix E provides a concise User's Manual of this template and should assist in enhancing one's overall understanding of the development and usage of this template.

2.6 Conclusion

This chapter discussed in detail the research on the development of automated Box Jenkins forecasting modeling and the implementation of this modeling using developed *Simphony* SPS template. The discussion initiated from the concise review of Box Jenkins forecasting methodologies. An examination regarding reasons for choosing Box Jenkins methodologies followed. Research ideas and different concerns and issues were introduced and explored thoroughly to provide necessary background knowledge in order to facilitate the understanding of the research.

In order to implement the proposed automatic Box Jenkins forecasting modeling, the core algorithms associated with the modeling are coded in *MATLAB* programming language. The implementation can, therefore, be realized in a computer environment. The integration of the automatic modeling is performed by the packaging and compiling elements of *MATLAB* programs. These elements employ core algorithms and a developed *Simphony* SPS template to facilitate the automatic execution of the Box Jenkins modeling process.

CHAPTER 3: RISK-BASED BOX JENKINS FORECASTING

3.1 Problems with Traditional Time Series Forecasting

Previous research has verified the applicability of basic forecasting methods, such as moving average, regression and exponential smoothing, in predicting future outcomes with a certain degree of accuracy. The predictions can serve as baseline information for decision-making or other detailed analysis. However, several researchers also noticed inevitable problems associated with forecasting methodologies and these problems have been identified and summarized in Chapter 1.

Based on the concerns of academic researchers on forecasting methodologies, it is obvious that a risk-based approach, which can work with time-series forecasting methodologies and perform risk and uncertainty analysis, should be proposed in order to overcome the problems inherent in the present system and to develop a more effective and powerful forecasting method.

3.2 Risk-based Box Jenkins Forecasting Approach

As mentioned in Chapter 1, the combination of time-series analysis and a simulation approach is still an area of development for engineering applications. Motivated by a lack of related research in the engineering area and by the need for the necessity of risk analysis-based forecasting methods, research was conducted with the aim of generating the specific methodologies for forecasting on capital planning using time-series simulation.

During this research, a wide range of literature on simulation technologies, Monte Carlo simulation methodologies, time-series process input modeling, and other related case studies were reviewed. The author attempted to summarize the contemporary research progress and to extract from it certain basic ideas for the development of algorithms and methodologies for engineering applications. However, considering the amount of available time, the research scope focused mainly on the study of input modeling methodologies and implementation methods.

Time-series simulation methodologies must have a time-series model as their basic numerical model, from which a simulation-based extension is created, in order to introduce the risk and uncertainty analyses. In this research, several similar studies were reviewed and finally Box Jenkins methodologies were determined to be the most suitable models for the required purpose.

3.2.1 Simulation Technologies in General

Before discussing time-series simulation technologies, the basic information about simulation as a general analytical tool should be covered.

Shannon (1975) defines simulation as "the process of designing a model of a real system and conducting experiments with this model for the purpose either of understanding the behaviour of the system or of evaluating various strategies (within the limits imposed by a criterion or set of criteria) for the operation of the system."

Simulation experiments are normally performed by computer. According to Page's PhD dissertation (1994), digital computer simulation may be divided into three categories: (1) Monte Carlo, (2) continuous, and (3) discrete event. Monte Carlo simulation is a method by which an inherently non-probabilistic problem is solved using a stochastic process. An explicit representation of time is not required. In a continuous simulation, the variables within the simulation are continuous functions, for example, a system of differential equations. If value changes to program variables occur at precise points in simulation time (that is, the variables are "piecewise linear"), the simulation is discrete event.

This thesis mainly deals with Monte Carlo simulation since this type of simulation is a suitable risk analysis tool in capital planning and budgeting for engineering and

management. The basic concept and methodologies of Monte Carlo simulation can be illustrated by the following descriptions.
Canada (1996) noted that Monte Carlo simulation "is an especially useful means of analyzing situations involving risk to obtain approximate answers when a physical experiment or the use of analytical approaches is either too burdensome or not feasible."

Monte Carlo simulation involves, first, the random selection of an outcome for each variable of interest; second, the combining of these outcomes with any fixed amounts; and, third, the calculation, if necessary, to obtain one trial outcome producing the desired answer. Repeated simulations will result in enough trial outcomes to obtain a sufficiently close approximation of the mean, variance, distribution shape, or other characteristic of the desired answer.

Figure 3-1, taken from Canada's (1996) *Capital Investment Analysis for Engineering and Management*, illustrates the process as it is applied to investment project analysis. Other details about Monte Carlo Simulation lie outside the scope of this thesis, which is concentrated primarily with the application of this technology in time-series simulation analysis. Interested readers are referred to *Project Management: Techniques in Planning and Controlling Construction Projects, 2nd Edition* (Ahuja et al. 1994) for detailed information.

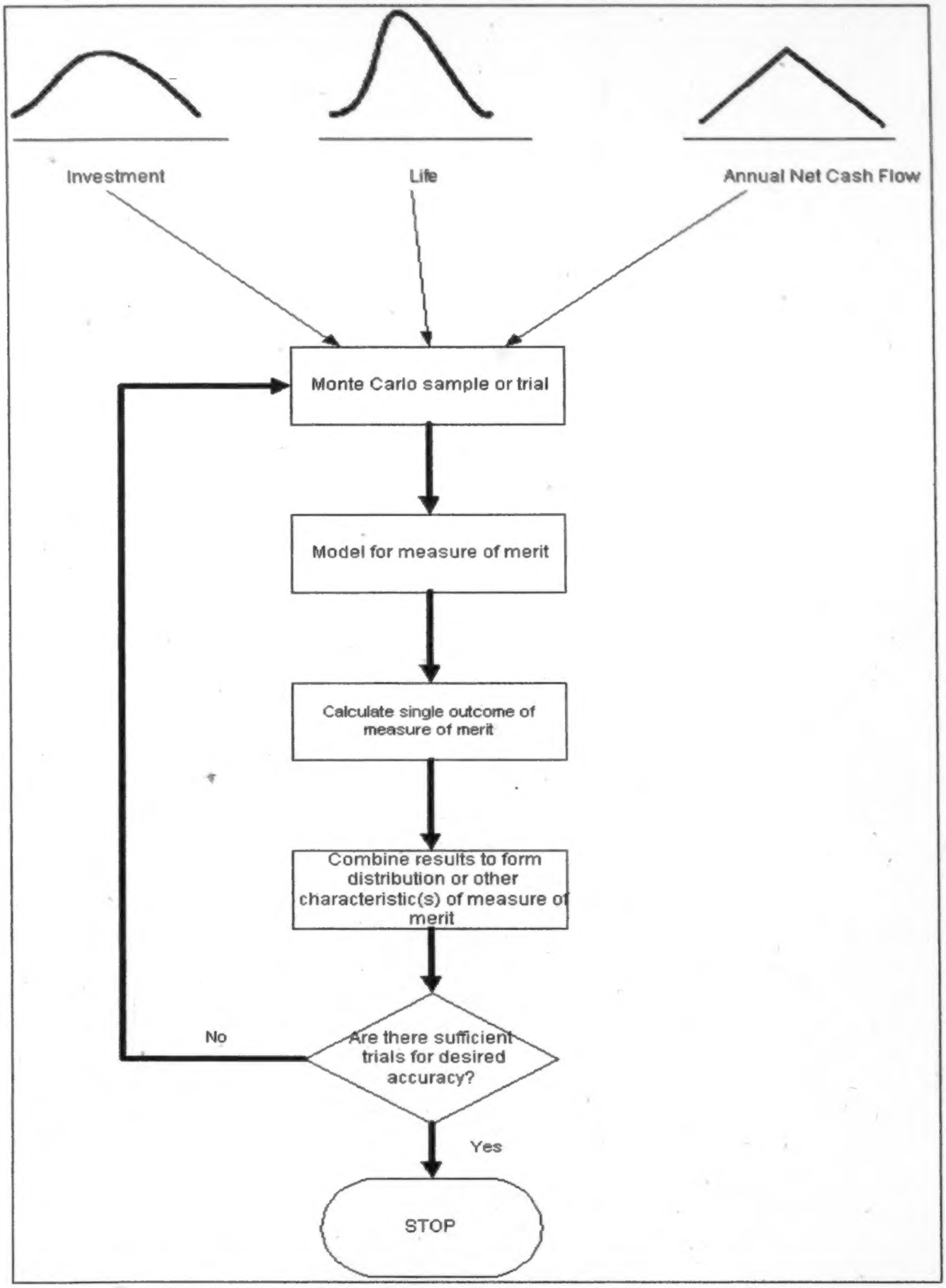

Investment
Life
Annual Net Cash Flow
Monte Carlo sample or trial
Model for measure of merit
Calculate single outcome of measure of merit
Combine results to form distribution or other characteristic(s) of measure of merit
Are there sufficient trials for desired accuracy?
No
Yes
STOP

Figure 3-1: Schematic Illustration of Monte Carlo Simulation Process – "Capital Investment Analysis for Engineering and Management", Canada (1996)

3.2.2 Concerns about Simulation Input Modeling

Normally, simulation analysis includes input modeling, simulation execution, and output analysis as the three major procedures, which comprise a complete analysis. Research will focus, accordingly, on the development of input modeling algorithms, on the development of simulation execution, or on the development of output analysis methodologies. This research attempted to develop and to test the specific input modeling methods for solving engineering problems, especially in the construction engineering and management area. The other two procedures employed a variety of established methodologies and research.

A simulation study will require an input modeling method to mimic the probabilistic nature of the system under consideration. A close match between the input model and the true underlying probabilistic mechanism associated with the system is necessary for successful input modeling (Leemis, 2001).

Appropriate input modeling proves to be very important to the success of a simulation experiment. Normally, input modeling involves the random selection of an outcome of variables, which are supposed to be risk factors. Nevertheless, simulators, frequently, assume that independent input variables are dominant so they do not consider the correlations between each random sampling when performing simulation calculations.

However, several researchers noticed the serious problems associated with an ignorance of dependences in simulation input modeling for special cases.

Biller and Nelson (2002) urged for the existence of dependent time-series input processes in the simulation of many service communication and manufacturing systems. They gave examples found in a continuous-flow production line and the sales of a large vehicle manufacturer. These data were recorded at fixed time intervals and their time-series plot revealed a strong dependence. Therefore, it was determined that an ignorance of these dependences in simulation could lead to performance measures that are seriously in error and towards a significant distortion within the simulated system.

Cario (1996) also identified several dependent processes in many real systems such as the temperature in a steel mill oven, as measured over time, the breadth of demands upon an inventory system, the daily reserve levels in a blood bank. These real systems are all likely to be autocorrelated. He noted that the prevalence of dependent input processes, as well as the potential error in the simulation output if dependences are ignored, will motivate the need ultimately, for simulation input models that represent autocorrelated time-series.

Leemis's (2001) review study on the use of input modeling techniques for discrete event simulation summarized the input modeling taxonomy demonstrated in Figure 3-2. This figure clearly demonstrates that input models can be classified into two branches according to their functionality in considering time dependency. The bottom branch can be further classified into several sub-branches or sub-sub-branches. Using this system of categorization, the author's attention will follow this progression:

Input Models→Stochastic Processes→Discrete-time→Continuous-state→Stationary or Not stationary→ARIMA (*p*, *q*) or ARIMA (*p*, *d*, *q*).

The Box Jenkins-based simulation approach is based on Box Jenkins forecasting methodologies. It employs ARIMA modeling to build a linear or non-linear relationship between current values and past values, or current values and past forecasting errors, or the combination of both. In this way, the appropriate dependency between different time points can be incorporated and future forecasts can be offered and sampled dependently.

The Box Jenkins based simulation approach is thought to be an effective approach to incorporating dependency considerations into input modeling. It accounts for the dependency between sampling variables along the timeline on discrete time points. Our research has verified that this approach can provide a comparably simple solution for performing time-dependent input modeling for Monte Carlo simulation.

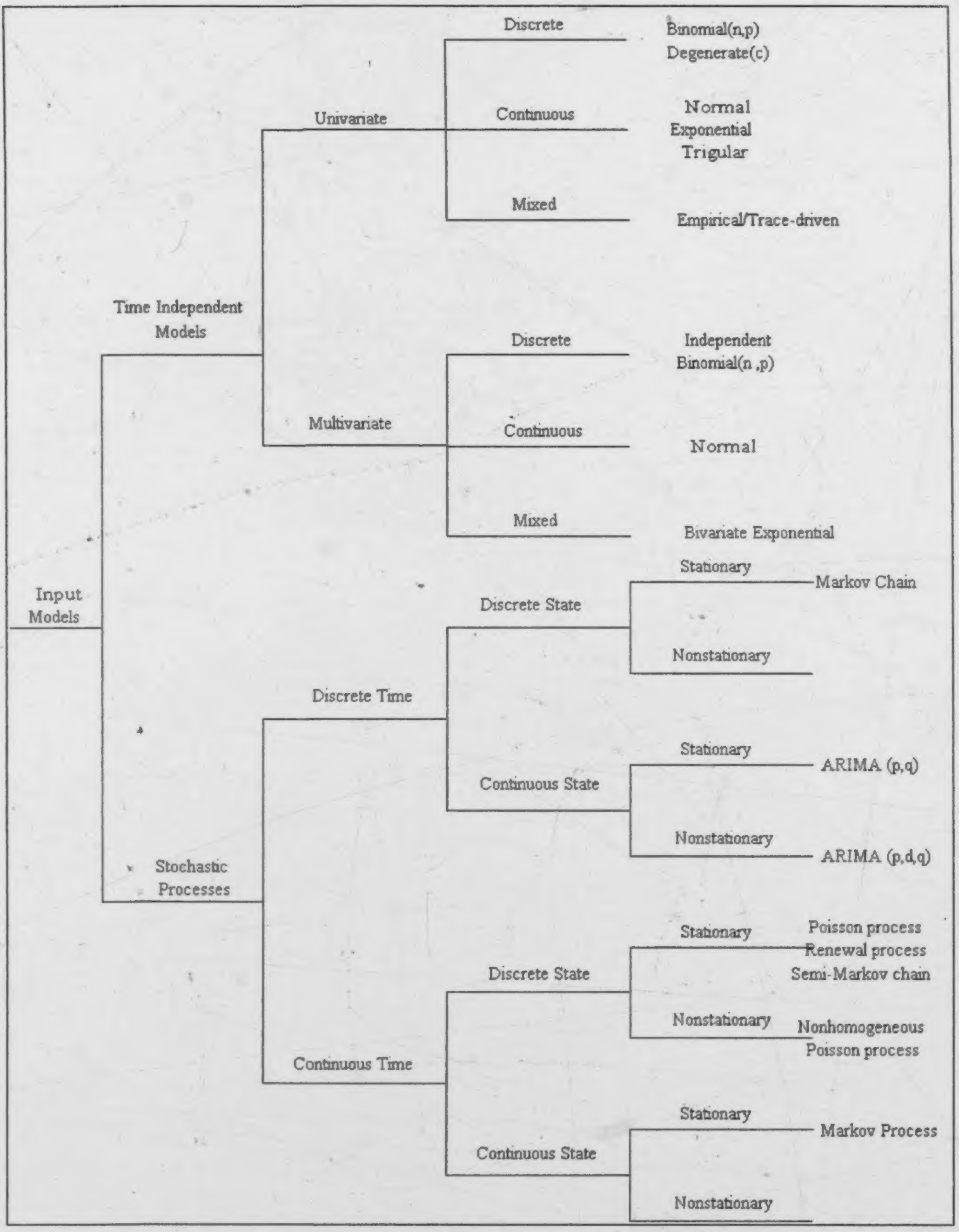

Input Models
Time Independent Models
Univariate
Discrete
Binomial(n,p)
Degenerate(c)
Continuous
Normal
Exponential
Trigular
Mixed
Empirical/Trace-driven
Multivariate
Discrete
Independent Binomial(n ,p)
Continuous
Normal
Mixed
Bivariate Exponential
Stochastic Processes
Discrete Time
Discrete State
Stationary
Markov Chain
Nonstationary
Continuous State
Stationary
ARIMA (p,q)
Nonstationary
ARIMA (p,d,q)
Continuous Time
Discrete State
Stationary
Poisson process
Renewal process
Semi-Markov chain
Nonstationary
Nonhomogeneous Poisson process
Continuous State
Stationary
Markov Process
Nonstationary

Figure 3-2: Taxonomy of Input Models (Leemis 2001)

3.2.3 Box Jenkins-based Monte Carlo Simulation Procedure

In the above paragraphs for the review of Box Jenkins methodologies, a generic ARIMA function was provided to represent the relationship between current values and past values, as well as with past forecasting errors.

$$Z_t = \delta + \phi_1 Z_{t-1} + \phi_2 Z_{t-2} + \phi_3 Z_{t-3} + \cdots + \varepsilon_t - \Theta_1 \varepsilon_{t-1} - \Theta_2 \varepsilon_{t-2} - \cdots \quad \text{(Equation 3-1)}$$

If we manipulate the above function and move the time point one unit ahead, we can get a linear function for the description of future values with current values and past values, as well as with future and past forecasting errors. This function provides us a mean with which to study the future outcomes.

$$Z_{t+1} = \delta + \phi_0 Z_t + \phi_1 Z_{t-1} + \phi_2 Z_{t-2} + \cdots + \varepsilon_{t+1} - \Theta_0 \varepsilon_t - \Theta_1 \varepsilon_{t-1} - \cdots \quad \text{(Equation 3-2)}$$

Equation 3-2 clearly has iterative characteristics. One may explain these characteristics by imaging that, with all the estimates of model parameters and current available values Z_t, Z_{t-1}, Z_{t-2}..., it is simple to estimate Z_{t+1}, which is the future value. If one assumes that, this forecast is the real future outcome. Moving the time point one unit ahead, another future value, Z_{t+2}, can also be obtained. If one continues to employ this iterative calculation, any length of Z_t can be obtained.

It should be stated that in the above deterministic forecasts, ε_t is assumed to be equal to zero. However, in simulation approach, ε_t must be considered as a randomly determined factor, which enable Monte Carlo simulation iterations to be executed. ε_t is normally assumed to be a normally distributed random variate and as such, it is sometimes called white noise residual. It has a predetermined mean of zero and a variance of one (Pankratz, 1983).

In Box Jenkins-based Monte Carlo simulation, several sample paths of simulated time-series data with predetermined lengths are generated and the mechanism of this generation operates using the following iterative procedures:

1. Move time point one unit ahead by transforming Equation 3-1 into Equation 3-2;
2. A pseudo-random number is generated that is uniformly distributed on the interval (0,1);
3. This random number is transformed into a normally distributed number with a mean of 0 and a variance of 1, by a transformation method such as Box-Muller of Inverse Method;
4. This normally distributed variate is deemed as a white noise error or residual and it is included in the ARIMA model (Equation 3-2) for the calculation of future values (other values are available as starting values);
5. Repeat Step 1 to 4 to generate other future values, assuming that the last estimated future value is the real observation, until all sampling values with length t have been generated;
6. Repeat Step 1 to 5 to continue with another iteration, until the predetermined iterations have been performed for Monte Carlo simulation analysis;
7. The output analysis can be performed on a large number of simulated future values on discrete future time points.

It is important for successful Monte Carlo simulation of the Box Jenkins time series model that the simulation calculation be based on the independent sample path. This requirement means that each sample path should be statistically independent in order to mimic the real system to the utmost degree. Therefore, the algorithm used for generating a random number is especially desired.

These basic algorithms are developed with a fitted ARIMA model using pre-sample data. Therefore, several assumptions and preliminary requirements must be established to enable the use of this approach:

1. The fitted ARIMA model is ready and all the estimates of model parameters have been obtained;
2. Original time-series data is processed as stationary and is ready to serve as starting values for simulation calculations;
3. The ARIMA model, together with those estimated parameters, is assumed to be comparably constant for extrapolation operations;
4. The system being studied is univariate and a fitted ARIMA model is adequate to represent the complex numerical relationship operating beneath the system.
5. The simulated future values basically follow a normal distribution when the number of simulation iteration or the number of simulated future values is large enough that a stable converge maybe obtained.

3.2.4 Box Jenkins-based Monte Carlo Simulation Methodologies

Considering the importance of simulation modeling methods and the inadequacy of related research on Box Jenkins-based Monte Carlo simulation, we developed a number of trial modeling methodologies. These methodologies follow the basic practices of simulation modeling procedures and consider the required concerns associated with the application. Certain methods came from the modification of similar work and improved upon the current methodologies.

AbouRizk and Halpin (1990) researched a number of simulation studies on repetitive construction processes. They proposed a complete set of algorithms and methodologies for simulation input modeling techniques and output analysis methods, including necessary verification and validation methods. They recommended several points, which the simulator should ensure for the success of the simulation experiment for the construction applications:

- Proper input in the form of statistical models;
- Proper analysis of the output; and

- Validation and verification of the model

Inspired and motivated by their research approach, this research gave special attention to the development of simulation modeling methodologies for Box Jenkins-based Monte Carlo simulation analysis. A complete set of methodologies are collected and summarized as follows, with a specific purpose to process time-series data using Box Jenkins technologies.

I. Data Collection

Proper data collection and related data processing methods are very important for successful time-series input modeling. Forecasting and simulation are only as accurate as the data used to generate summary statistics and forecasts. The most sophisticated statistical techniques and forecasting model will be useless if they are applied to unreliable data.

Normally, for successful input modeling, Box Jenkins methodologies require 50 to 100 data points to build a good model. Box and Jenkins (1976) suggested that about 50 observations is the minimum required number of data points. Other analysts mentioned that if using a smaller sample size, an interpretation of the results should be made with caution. A large sample data set is desirable particularly when one is working with seasonal data.

Three traditional types of problems may come with collected data. Therefore, certain methods should be employed to process original data into modeling data acceptable to Box Jenkins methodologies.

Missing data: may be caused by missing observations in a data series. Another kind of missing data is caused by the subjective deletion of outlets. Several methods can be applied in dealing with missing data. The following list contains four basic methods:

- Replace missing data with the mean of a series;
- Replace missing data with a naïve forecast;
- Replace missing data with a simple trend forecast, or
- Replace missing data with an average of the last two known observations, which bound the missing observations.

These methods have their own degrees of applicability and their own limitations. In most cases, the fourth method is the easiest and most feasible method to employ.

Existence of outlets: should be removed from the original data set before any further processing. These irregular data points are caused by the influence of outside events, which are unpredictable. The existence of such outlets will damage the original data and cause impact on the further analysis. It should be noted that the removal of these outlets also results in missing data.

Data frequency problem: A specific problem associated with Box Jenkins-based simulation analysis is frequency. Often, time-series data are not presented in a manner conducive to the research being undertaken. Box Jenkins methodologies, however, require time-series data collection with high frequencies. This conflict requires the application of one of the following frequency conversion methods.

- **Repetition**
- **The equal-step method**
- **Linear or non-linear growth**

The equal-step method is the most simple of the three and most capable of capturing the upturn or downturn in time-series data.

All these problems can be identified using graphical presentation methods. These methods are frequently used throughout the whole process of input modeling for the observation of time-series data, for the graphical presentation of statistical analysis, and for the performance of a hypothesis test. However, they also offer the most challenges for simulators when human judgements are needed.

- **Single-scale plot of a time-series**:
- **Scatter Diagram**
- **Bar Graphs**

The detailed discussions on the above methodologies are beyond the limit of this thesis. Interested readers are referred to literatur on time-series forecasting methodologies for related information.

II. Data Stationarity Transformation

Original time series data, which are collected in a proper way, that is, which undergo necessary processing if required, constitute the basis for further analysis, may go through several transformation procedures in order to remove certain specific components, which may do harm to the input modeling. Basically, Box Jenkins-based input modeling requires continually sequenced, statistically stationary, and high frequency time-series data. However, in natural situations, these kinds of data are rare; original data, therefore, need to be processed in order to achieve these standards.

This achievement requires the identification of several characteristics pertaining to time-series data:

- **Trend**: a persistent upward or downward movement of data over a long period of time. This component demonstrates that data are not stationary in mean value.
- **Seasonal Variations**: a pattern of change in the data, which completes itself within a calendar year and is then repeated on a yearly basis. This component shows that data are influenced by certain seasonal factors.

- **Cycle**: the upward and downward change in the data pattern that occurs over a duration of two to ten years or longer. Cycle is one of the most difficult characteristics to forecast because of its longer time frame.
- **Irregular or error fluctuations**: sometime called "white noise" or "outlets". This characteristic should be removed from the original data set. Otherwise, related analysis will be extremely inaccurate.

Normally original time-series data can contain a combination of these four characteristics, which demonstrates the complex variation along the time axis. A simple solution for identifying these characteristics is to observe the data visually through the graphical presentations methods mentioned in the above section.

After the observation of a characteristics or combination of these characteristics, a series of transformation methods can be applied to remove these components. To clearly illustrate these methods, the concepts of stationarity and the several basic types of unstationarity should be defined.

Stationarity means that there is no growth or decline in the data. The data must remain roughly horizontal along the time axis. In other words, the data fluctuate around a constant mean, independent of time, and the variance of the fluctuation remains essentially constant over time. Three kinds of unstationarity, according to the description of stationarity, exist with original time-series data. They are:

- Unstationarity in the mean

This kind of unstationarity exists when the time-series plot shows a change in the mean over time.

- Unstationarity with variance

When the time-series plot demonstrates an obvious change in the variance over time, it refers to unstationarity in the variance.

- Unstationarity with seasonality

If the time series plot shows seasonal up and down over time, it is called seasonal unstationarity.

Normally, real time series data contain one or more of the above unstationarity, thereby affecting the plot of time-series data in a simple or complex way. Therefore, it is difficult to identify the unstationarity of data from the very beginning.

Several transformation methods have proven to be effective in offsetting these characteristics. They can be used independently or together for addressing single or combined unstationarity components.

- **N-order Normal Differencing**
- **Seasonal Differencing**
- **Natural logarithm transformation**

Much literature has introduced methods of judging the stationarity of time-series data. These methods can be classified into two branches: graphical methods and numerical methods. Due to the scope of this thesis, general summaries about these methods should suffice as explanations and these are provided along with discussion of an automatic transformation method.

- **Graphical judgement methods**

Simple plotting of time-series data, whether original or processed, is a direct and quick way to examine the stationarity. A certain trend or seasonality can be identified through visual observation. Related transformation methods can then be applied.

ACF and PACF are two measurements for the examination of correlations within the time-series data. Barcharts based on their measurements over a lag-time should be used on the original data and processed data after each processing is done.

- **Numerical judgement methods**

Sometimes it is difficult to identify graphically the stationarity of data based on a nonstationary time-series, therefore, more accurate numerical methods should be employed.

Currently available numerical methods apply a hypothesis test as an efficient way to examine the stationarity. In theory, if a time-series is stationary without seasonality, the autocorrelations (ACF and PACF) should all be zero or else they will differ from zero only for the first few lags. For large samples, in order to test whether the autocorrelations are statistically equal to zero, one should use *t*-statistic, associated with ACF or PACF and compare them with predetermined critical values. In this way, the stationarity of time-series data can be judged numerically.

In theory, *t*-test for ACF and PACF must be rigidly obeyed in order to assure the stationarity of time-series data. However, in reality, as the number of sample data increases, the probability of rejecting at least one true null also increases. Thus to facilitate the testing, the condition is usually relaxed in order to allow one or two t-statistics in the first lag or first two lags to be larger than the acceptable critical level. This compromise proves to be feasible and practical in real modeling practices.

The objective of data transformation is quite straightforward. Stationary time series data is obtained in order to meet the assumptions of Box Jenkins methodologies. Time-series data should be statistically stationary in mean and variance, without seasonality and trend, in order for different transformation methods to be applied according to both graphical and numerical judgements. Modeling experience proves that a certain application sequence should be followed in order to obtain the stationarity quickly. Basically, the sequence follows Natural Logarithm Transformation first, if necessary, then seasonal differencing, if necessary, and normal differencing.

III. ARIMA Model Fitting

ARIMA Model fitting for Box Jenkins-based Monte Carlo simulation can be described as the identification of a fitted ARIMA model and the estimation of model parameters. Therefore, the methodologies for model fitting share the most similarity with Best ARIMA Model Idenfication for Box Jenkins forecasting modeling.

All the methodologies for can be referred from literature about Box Jenkins forecasting modeling. Since ARIMA model fitting is also a time consuming and iterative procedure, the proposed automated Box Jenkins forecasting modeling tool in Chapter 2 applies as well in facilitating ARIMA model fitting. In the following section about implementation of Box Jenkins-based Monte Carlo simulation, readers will find this tool is utilized to perform input modeling.

IV. Input Modeling Finalization

Once these procedures have been completed and the best fitted ARIMA model is obtained, it is necessary to finalize the input modeling procedures and prepare for the Monte Carlo simulation.

The following results should be collected and recorded:

- Best fitted ARIMA model, along with all estimated parameters;
- Stationary time-series data based on the transformation of original-time series data;
- Original time-series data;
- Final order of normal differencing;
- Final lag of seasonal differencing; and
- Binary information regarding whether or not natural logarithm transformation has been performed.

3.2.5 Box Jenkins-based Monte Carlo Simulation Execution and Output Analysis

After input modeling is finished, the resulting ARIMA model and related data can be fed into the Monte Carlo simulator for iterative execution.

Although the iterative execution is not complex, it is time-consuming. Basically, simulators must specify the number of iterations and the number of periods of time to extrapolate based on a current position in time.

The following figure illustrates the Box Jenkins-based Monte Carlo simulation execution:

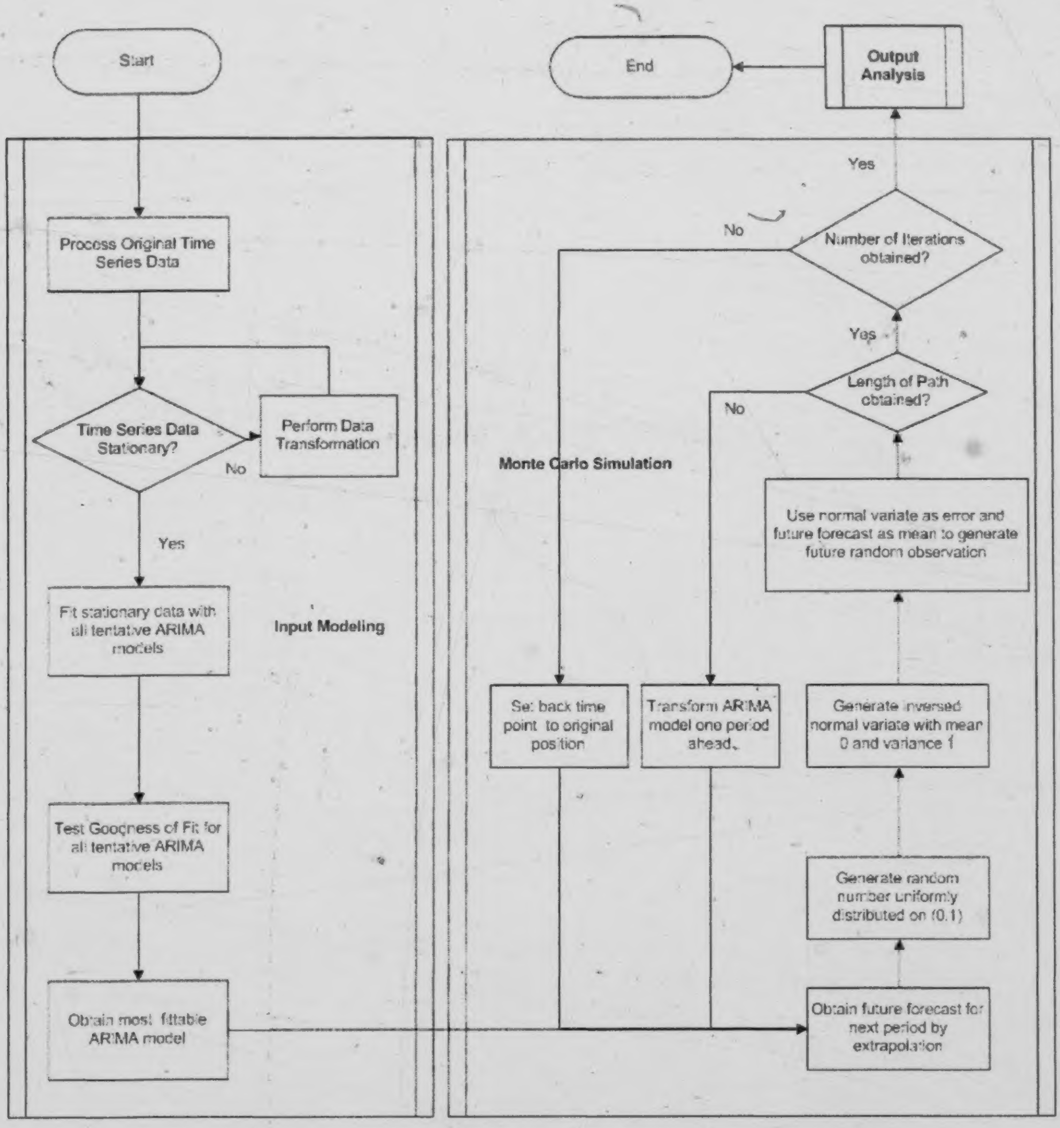

Start
End
Output Analysis
Yes
No
Number of Iterations obtained?
Process Original Time Series Data
Yes
Length of Path obtained?
No
Time Series Data Stationary?
Perform Data Transformation
No
Monte Carlo Simulation
Use normal variate as error and future forecast as mean to generate future random observation
Yes
Fit stationary data with all tentative ARIMA models
Input Modeling
Set back time point to original position
Transform ARIMA model one period ahead
Generate inversed normal variate with mean 0 and variance 1
Test Goodness of Fit for all tentative ARIMA models
Generate random number uniformly distributed on (0,1)
Obtain most fittable ARIMA model
Obtain future forecast for next period by extrapolation

Figure 3-3: Box Jenkins-based Monte Carlo Simulation Algorithms

The results from Box Jenkins-based Monte Carlo simulation consist of a huge matrix of simulated future variates in which the number of random paths of observations corresponds to the row amount, while the number of periods to extrapolate corresponds to the column amount. All the results should then be inversely processed based on stationary time-series data into unstationary time-series data, based on the transformation information recorded in the input modeling finalization procedure. Three related inverse processing methods are identified:

- Unnormal differencing;
- Unseasonal differencing; and
- Unnatural logarithm transformation.

The algorithms of these methods are easy to understand and to implement if one knows how to perform the original transformations.

In theory, a large number of iterations are recommended because the Monte Carlo simulator requires a large number of replications to be generated, from which empirical distribution of extrapolated sampling data points can be created in order to approximate an actual distribution.

The number of periods should be selected with extreme caution. In Chapter 2, it was mentioned that Box Jenkins methodologies, as foundational to Box Jenkins-based Monte Carlo Simulation, are more reliable in short or medium-range forecasting. Certain researchers, therefore, recommended that no more than five periods would constitute a reliable selection if the ARIMA model is perfectly fitted (Gaynor and Kirkpatrick 1994).

It should be clarified, however, that even with the limitation in long range forecasting, Box Jenkins-based Monte Carlo simulations still behave better than the simple forecasting approach. The traditional forecasting approach only provides point estimates or range estimates with a certain degree of confidence level. The Box Jenkins-based approach, however, is able to provide stochastic analysis with more statistical information and more risk analysis results. In terms of available and pertinent information on the uncertainty in the future, Box Jenkins-based Monte Carlo simulation is considerably more helpful than the traditional approach.

After a large number of replications are obtained for each future period, the output analysis for the simulation experiment can be performed. As mentioned before, we believe that the simulated variates from Box Jenkins-based Monte Carlo simulation should follow a normal distribution because these variates are actually the summation of deterministic forecasts and randomly generated white noise or forecasting error, which is believed to follow a standard normal distribution. With this in mind, we are, therefore, able to employ established research findings on simulation output analysis in order to perform this analysis.

AbouRizk and Halpin's research on "Probabilistic Simulation Studies for Repetitive Construction Processes" (1990) introduced methods of performing output analysis for simulation experiments where the resulting random variates are normally distributed.

Estimating Mean

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \text{ -- (Equation 3-17)}$$

Where X is the simulated variate and n is the number of simulated variates.

Estimating Variance

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 \text{ -- (Equation 3-18)}$$

Where S^2 is the variance and n is the number of simulated variates.

Estimating Arbitrary Quantitile Points

$$X_q = \overline{X} + z_q S \text{ -- (Equation 3-19)}$$

where X_q is the qth quantitle point of the random variates, z_q is the critical value from the standard normal distribution at the specified cut-off value q, and S is the MLE of the sample standard deviation, which is obtained from the square of S^2 (from Eq. 3-18) multiplied by *(n-1)/n*.

Estimating Probabilities

$$\Pr(X \leq x) = \Phi(\frac{x-\overline{X}}{S}) \text{--- (Equation 3-20)}$$

Where $\Pr(X \leq x)$ is the probability that the output X does not exceed a particular fixed value of x.

Other classic output analysis methods like Cumulative Distribution Function (CDF) and Probability Distribution Function (PDF) and the methodologies for descriptive statistics can be easily referred to in the many articles on Engineering Statistics.

3.3 Risk-based Box Jenkins Forecasting Implementation

Box Jenkins-based Monte Carlo simulation is actually an extension of Box Jenkins forecasting functionalities. The implementation of Box Jenkins-based Monte Carlo simulation, therefore, shares certain techniques with the implementation of Box Jenkins forecasting.

As mentioned in Chapter 3, Box Jenkins forecasting outputs, that is, the fitted ARIMA model and various transformation parameters, together with the original time-series data and stationary time-series data become the input model for the Monte Carlo simulation experiment. Thus, the implementation of a Box Jenkins-based Monte Carlo simulation can actually use the same tool associated with automated Box Jenkins forecasting modeling for its input modeling.

An implementation of Box Jenkins-based Monte Carlo simulation will include three procedures:

- Input modeling;
- Monte Carlo simulation execution; and

- Output analysis.

Methods or algorithms, related to this implementation, except those already mentioned in Box Jenkins forecasting modeling, include:

- Box Jenkins-based Monte Carlo sampling;
- Empirical CDF plotting;
- Histogram plotting;
- Descriptive statistics analysis;
- Percentile points estimation; and
- Probability estimation.

Section 3.2.3 gives a thorough introduction of Box Jenkins-based Monte Carlo sampling. This particular method is implemented by a computer program coded in MATLAB programming language. For reference purposes, the completed computer codes are included in Appendix C.

Other methods of implementation, such as output analysis methods, are similar to the basic methods for general simulation experiments. In Section 3.2.5, there was an extensive discussion of the theories underlying these methods. Based on this information, the implementation can be realized in a spreadsheet program, such as Microsoft Excel.

As mentioned above, the developed automatic Box Jenkins forecasting modeling tool can be used as an input-modeling tool for Box Jenkins-based simulation. In a simulation experiment, however, the simulator should pay more attention to the fitted ARIMA model than to the Box Jenkins-based forecasts. In addition, information regarding the processing of original time-series, such as the order of normal differencing or the seasonal lag of seasonal differencing and the processing of the resulting stationary time-series data will become components of the input modeling data.

3.4 Conclusion

This chapter discussed in detail the research on Risk-based Box Jenkins forecasting, that is, Box Jenkins based Monte Carlo simulation. The research included a summary of problems regarding traditional time-series forecasting in the first place. Then to solve these problems, a risk-based Box Jenkins forecasting approach was proposed by the introduction of background knowledge for reference purpose.

Box Jenkins-based Monte Carlo simulation can be implemented by employing the developed SPS template as an input-modeling tool, along with the designed MATLAB programs for the simulation execution. All related output analysis could be conducted by spreadsheet software such as Microsoft Excel, since the detailed methods and related equations are well established by past research work.

CHAPTER 4: MODEL VALIDATION AND CASE STUDY

4.1 Model Validation

Model validation proves to be an important task in defending and verifying methodologies. Any qualitative statement regarding the advantages or strengths of a methodology is not as convincing as a scientific validation test using data collected from an actual situation.

To verify that Box Jenkins forecasting methodologies are superior to other methods and to test the developed automatic modeling tool, a validation test for Box Jenkins forecasting methodologies was performed with the assistance of a developed *Simphony* SPS template. Another validation test was used to verify the functionality of Box Jenkins-based Monte Carlo simulation in risk and uncertainty analysis and to test the extension of forecasting functionalities. These two tests use testing data from public data resources to perform an analysis of the capital planning forecasting task. The basic validation methodology is to perform a comparative analysis between traditional methodologies and our proposed methodologies; the validation conclusions can then draw upon this comparison in terms of accuracy, functionality, and usability.

4.1.1 Introduction

In order to validate Box Jenkins methodologies using economic or technical data associated with the capital planning of infrastructure system, the testing data are carefully chosen within the domain of capital investment analysis. The data is selected to provide statistics for either the provincial or federal government level. Several criteria are identified:

1. The data are original, without artificial manipulation or processing;
2. The data are construction engineering and management related;

3. The overall summary of the data occurs at a high level of management;
4. The data are frequently collected with an identical interval; and
5. There are more than 50 data points collected, in total.

Criteria 1, 4 and 5 are considered based on the perspective of a statistical modeling algorithm. Criterion 5 is especially recommended by Box Jenkins methodologies. Other criteria are chosen because the author's focus is on construction engineering and management at high levels. The data that meet these criteria are highly suitable for related analyses.

Finally, a whole set of data was identified and selected from the website of the U.S. Census Bureau, Construction Section (http://www.census.gov/const/www/). Based on the above considerations, the author selected data from the U. S. Federal Construction Spending on Non-residential Projects, which focused primarily on infrastructure system spending. These data were collected monthly from January 1993 until July 2003 without seasonal adjustments. The total number of data points is 127 and the basic unit of the data is millions of dollars. All these data are tabulated and plotted in Table 4-1 and Figure 4-1.

Table 4-1: Tabulated Validation Data from US Census Bureau

Monthly Federal Construction Spending from Jan 1993 to July 2003 (Millions of Dollars)

Year	Total	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
1993	14,424	1,043	1,032	1,366	1,319	1,059	1,174	1,278	1,196	1,704	962	1,132	1,159
1994	14,440	1,009	1,066	1,019	1,103	1,108	1,252	1,173	1,242	1,712	1,112	1,237	1,407
1995	15,751	1,183	1,109	1,393	1,224	1,353	1,357	1,301	1,395	1,985	871	1,302	1,278
1996	15,325	1,193	1,132	1,242	1,304	1,416	1,314	1,283	1,261	2,005	924	1,188	1,062
1997	14,087	1,045	1,004	1,097	1,063	1,136	1,148	1,163	1,181	1,741	979	1,245	1,284
1998	14,318	1,006	1,035	1,134	1,069	1,081	1,197	1,347	1,233	1,823	921	1,243	1,230
1999	14,025	990	1,129	1,144	1,152	1,033	1,185	1,208	1,169	1,584	971	1,235	1,226
2000	14,166	976	1,077	1,254	1,094	1,109	1,248	1,061	1,291	1,692	1,001	1,205	1,158
2001	15,081	1,172	1,121	1,227	1,183	1,247	1,302	1,200	1,340	1,814	1,073	1,188	1,213
2002	16,313	1,144	1,264	1,313	1,283	1,373	1,353	1,364	1,404	1,901	1,188	1,354	1,373
2003	#N/A	1,276	1,261	1,279	1,434	1,466	1,535	1,449					

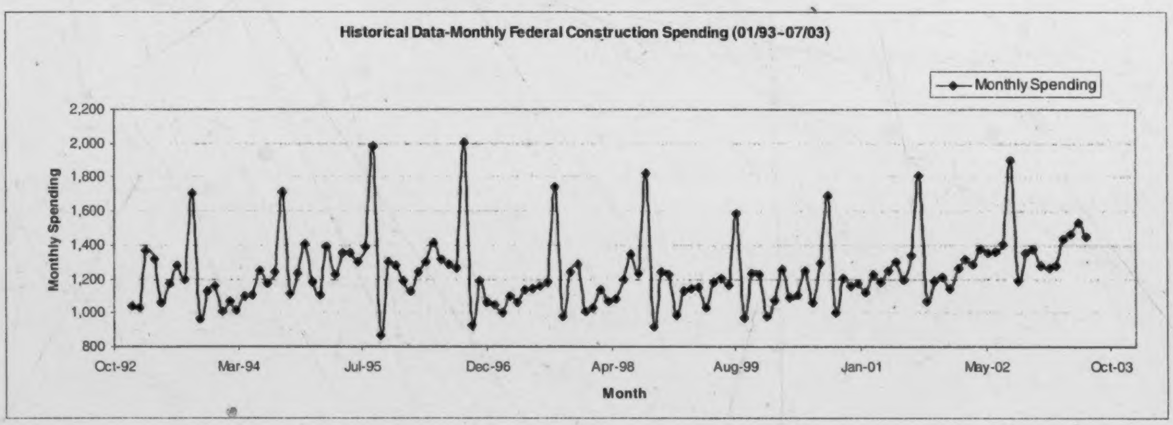


Figure 4-1: Time-series Plot of Validation Data from US Census Bureau

4.1.2 Validation of Automated Box Jenkins Forecasting

In order to validate the forecasting functionality, the author began by generating a time-series based on the original data set used for testing. The monthly data from January 1993 to July 2003 were organized as time-series data, situating the most recent data set collected (July 2003) as the last data point and the oldest data collected (Jan 1993) as the first point. In this way, an original time-series can be generated.

Based on the graphical observation of original time-series data, one identified a strong seasonality that appeared to control, to a certain degree, the fluctuation of monthly spending along a time axis. A slight upward trend is also present which may point to a continuous increase in the federal construction investment on non-residential projects over subsequent years. No irregular data points were identified other than those involved in the upward trend; therefore, the whole data set is suitable for further analysis.

Following similar practices of model validation, one need to choose part of the original data as a testing set and leave the remainder for modeling purposes. Considering the aforementioned characteristics of the original data, one decided to choose data points from September 2002 to July 2003 (a total of 11 data points) as the testing data. The remaining 116 data points will be used for forecasting modeling. For time-series data, it should be noted that the selection of testing data must not damage the continuity of a time-series; testing data, therefore, should be continuously selected from the most recent period.

The objective of this project is to verify the forecasting functionality of Box Jenkins methodologies through a comparison of its forecasts with those provided by other forecasting methods. Three forecasting methods are selected for the comparison: Moving Average, Regression, and Exponential Smoothing. These three methods are currently the most popular numerical methods available for forecasting and estimation in capital planning. These methods are easier to use and more understandable theoretically than Box Jenkins methodologies. Two graphical representations comparing the forecasting performance of all four methods are provided to summarize the test. Table 4-2 tabulates the original testing data against the forecasts for each method and; MAPE (Mean of Absolute Percentage Error) is tabulated as the key comparison criteria. Figure 4-2 is the graphical representation of the comparison of forecasts for all methods using original data as a baseline.

From these representations, one can easily observe the difference between four forecasting methods in terms of forecasting accuracy and fitting quality.
Other than the method using Box Jenkins methodologies, the methods all fail to grasp the complex variation of an original time-series, especially the seasonal fluctuation cycle that occurs each September.

In terms of forecasting errors level, the regression method gives the worst performance. Exponential smoothing and moving average methods provide similar error levels. By contrast, the Box Jenkins method offers the best forecasting accuracy and fitting quality. It excels in grasping the seasonal fluctuation mentioned above as well as in short-range forecasting (1~5 forecasting periods). The results of this method follow nearly the same variation pattern as the original time-series. For long-range forecasting as well, it provides the most accurate forecasts available for monthly spending from February 2003 to July 2003.

More modeling details can be found in Appendix A. An observation of the comparison results yielded the following conclusions:

- The moving average method fails because it is not capable of analyzing seasonality. Compared with the complexity of original time-series, its algorithm seems too simple;
- The exponential smoothing method relies heavily on former forecasts. This characteristic prohibits it from providing fast, long-range forecasts;
- The regression method proves to be unsuccessful in dealing with univariate time-series data. Since no data for independent factors can be identified, only time is regressed against the original data.

In general, this validation test shows that Box Jenkins methodologies have the best performance when dealing with complicated, seasonal, and non-linear time-series data. Box Jenkins methodologies can provide excellent short and medium-range forecasts. Long range forecasting, is also more effective than with other basic methods, especially in terms of accuracy and quality of data fitting.

L

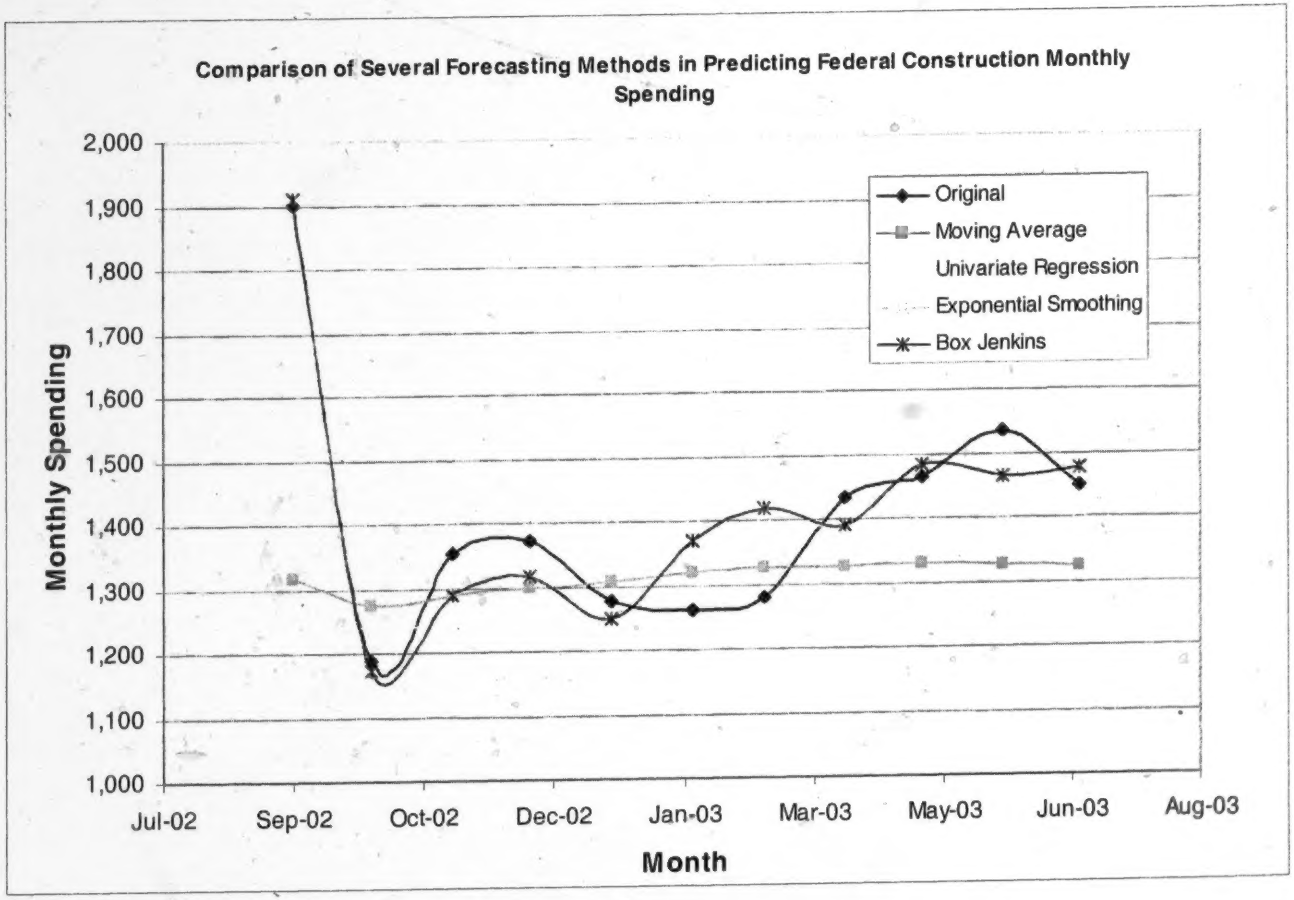


Figure 4-2: Comparison of Four Forecasting Methods in Terms of Accuracy

Comparison of Four Forecasting Methods										
Month	Original	Moving Average		Regression		Exponential Smoothing		Box Jenkins		
		Forecasts	APE	Forecasts	APE	Forecasts	APE	Forecasts	APE	
Sep-02	1,901	1,316	[illegible]	1239.053643	34.82%	1390.8363	26.84%	1912.0447	0.58%
Oct-02	1,188	1,274	7.24%	1239.291858	4.32%	1390.8363	17.07%	1172.5099	1.31%
Nov-02	1,354	1,291	4.67%	1239.530072	8.46%	1390.8363	2.72%	1288.9751	4.80%
Dec-02	1,373	1,299	[illegible]	1239.768286	9.70%	1390.8363	1.30%	1315.4403	4.19%
Jan-03	1,276	1,306	2.29%	1240.006501	2.82%	1390.8363	9.00%	1247.9054	2.20%
Feb-03	1,261	1,320	4.68%	1240.244715	1.65%	1390.8363	10.30%	1369.3706	8.59%
Mar-03	1,279	1,325	3.57%	1240.482929	3.01%	1390.8363	8.74%	1419.8358	11.01%
Apr-03	1,434	1,326	7.55%	1240.721144	13.48%	1390.8363	3.01%	1391.301	2.98%
May-03	1,466	1,329	[illegible]	1240.959358	15.35%	1390.8363	5.13%	1482.7662	1.14%
Jun-03	1,535	1,326	13.65%	1241.197572	19.14%	1390.8363	9.39%	1464.2313	4.61%
Jul-03	1,449	1,323	8.68%	1241.435787	14.33%	1390.8363	4.01%	1476.6965	1.91%
MAPE			8.90%		11.55%		8.86%		3.94%

Table 4-2: Comparison of Four Forecasting Methods In Terms of Accuracy

4.1.3 Validation of Risk-based Box Jenkins Forecasting

Box Jenkins-based simulation is especially advantageous in its ability to provide risk and uncertainty analysis to traditional forecasting. Its ability to incorporate dependency in random variable sampling also transcends traditional sampling techniques. The validation test aims to verify these two strengths through an implementation of functionalities.

4.1.3.1 Test on Risk and Uncertainty Analysis Functionality

This research demonstrates that Box Jenkins methodologies not only provide excellent forecasting accuracy in short and medium range forecasting, but can also be extended to incorporate a risk and uncertainty analysis through the use of simulation analysis. Simulation analysis can be used to obtain rich and comprehensive statistics; this information will improve our confidence in the future variations.

A standard Box Jenkins-based Monte Carlo simulation was performed using the same testing data used for testing the forecasting functionality. The results were summarized, tabulated, and compared with the results gathered from Box Jenkins forecasting. This method allowed us to observe difference between Box Jenkins forecasting and Box Jenkins-based simulation in terms of the availability of rich statistics, which are required for risk and uncertainty analysis.

Figure 4-3 shows the results of 100 simulated observation paths along a time axis with time points situated between September 2002 and July 2003. These simulated observations are used for risk and uncertainty analysis. The related methodologies and implementation methods were discussed in Chapter 3 and the previous sections.

Figure 4-4 shows the empirical CDF and Histogram for simulated observations of four example time points (September 2002 to December 2002). Descriptive statistics were collected for the four example time points and associated Rank and Percent Analysis was performed. These results were tabulated in Table 4-3 and Table 4-4.

All the above analyses constitute a basic Monte Carlo simulation Output Analysis. For demonstration purposes, only 100 iterations of the calculations were executed and all analyses were performed on the four example time points. A complete simulation experiment could require a higher number of iterations to be executed and the simulation output analysis would have to be done on all related future periods.

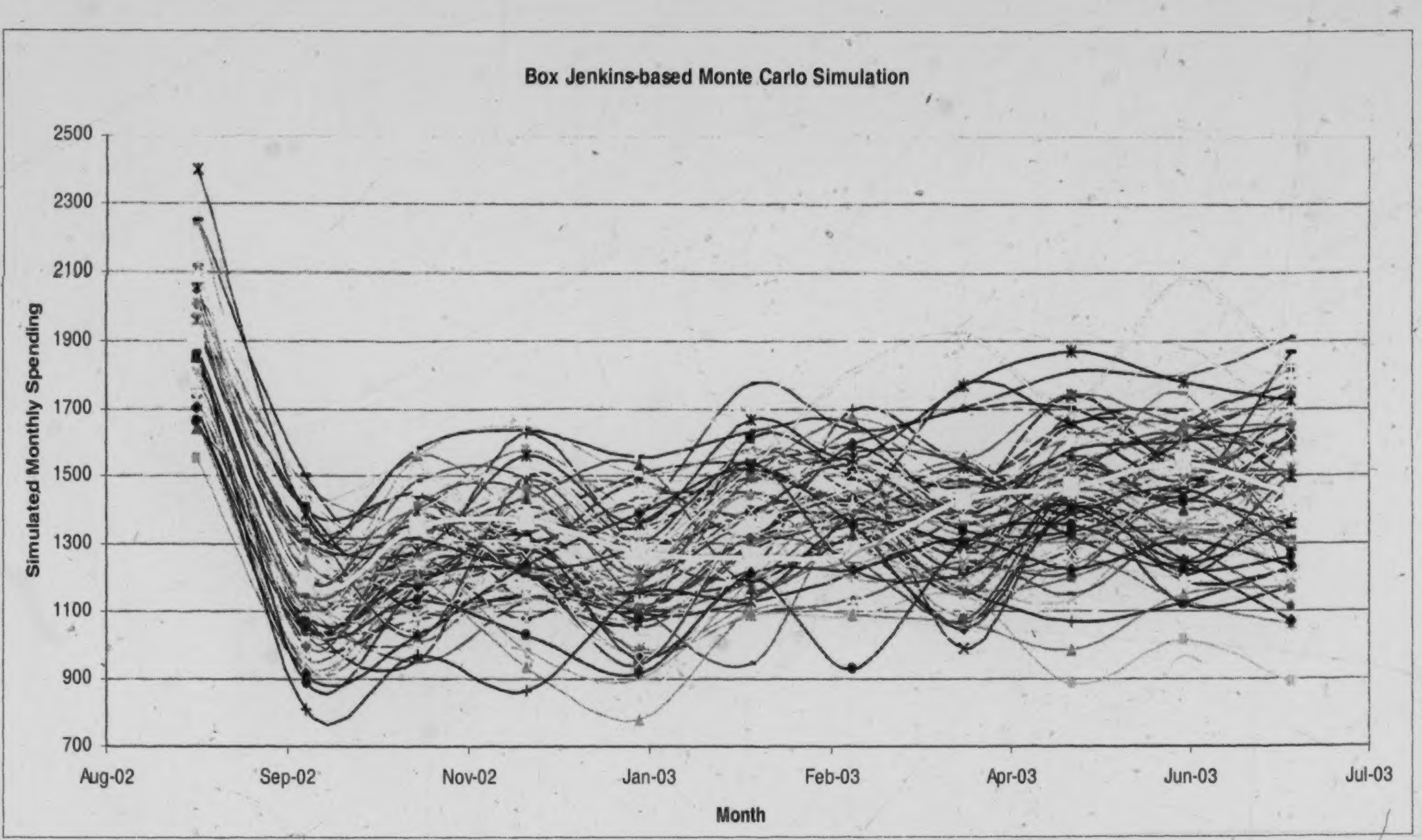


Figure 4-3: Time-series Plot for 100 Simulation Paths

Descriptive Statistical Analysis (Box Jenkins Methodologies)

Sep-02		Oct-02		Nov-02		Dec-02	
Original	1901000	Original	1185000	Original	1354000	Original	1373000
Mean	1917.9407	*Mean*	1185.5325	*Mean*	1265.4054	*Mean*	1290.4273
Standard Error.	15.758964	*Standard Error*	14.936416	*Standard Error*	15.245037	*Standard Error*	17.317154
Median	1930.1176	*Median*	1176.7401	*Median*	1266.9581	*Median*	1272.8412
Mode	#N/A	*Mode*	#N/A	*Mode*	#N/A	*Mode*	#N/A
Standard Deviation	157.58964	*Standard Deviation*	149.36416	*Standard Deviation*	152.45037	*Standard Deviation*	173.17154
Sample Variance	24834.496	*Sample Variance*	22309.651	*Sample Variance*	23241.117	*Sample Variance*	29988.383
Kurtosis	0.1909275	*Kurtosis*	-0.518866	*Kurtosis*	-0.226686	*Kurtosis*	-0.203338
Skewness	-0.046192	*Skewness*	-0.038403	*Skewness*	0.1006718	*Skewness*	0.1164927
Range	849.039	*Range*	692.69593	*Range*	710.80433	*Range*	811.18355
Minimum	1550.1458	*Minimum*	810.84117	*Minimum*	949.91297	*Minimum*	863.18515
Maximum	2399.1848	*Maximum*	1503.5371	*Maximum*	1660.7173	*Maximum*	1674.3687
Sum	191794.07	*Sum*	118553.25	*Sum*	126540.54	*Sum*	129042.73
Count	100	*Count*	100	*Count*	100	*Count*	100

Table 4-3: Descriptive Statistics for Four Example Time Points

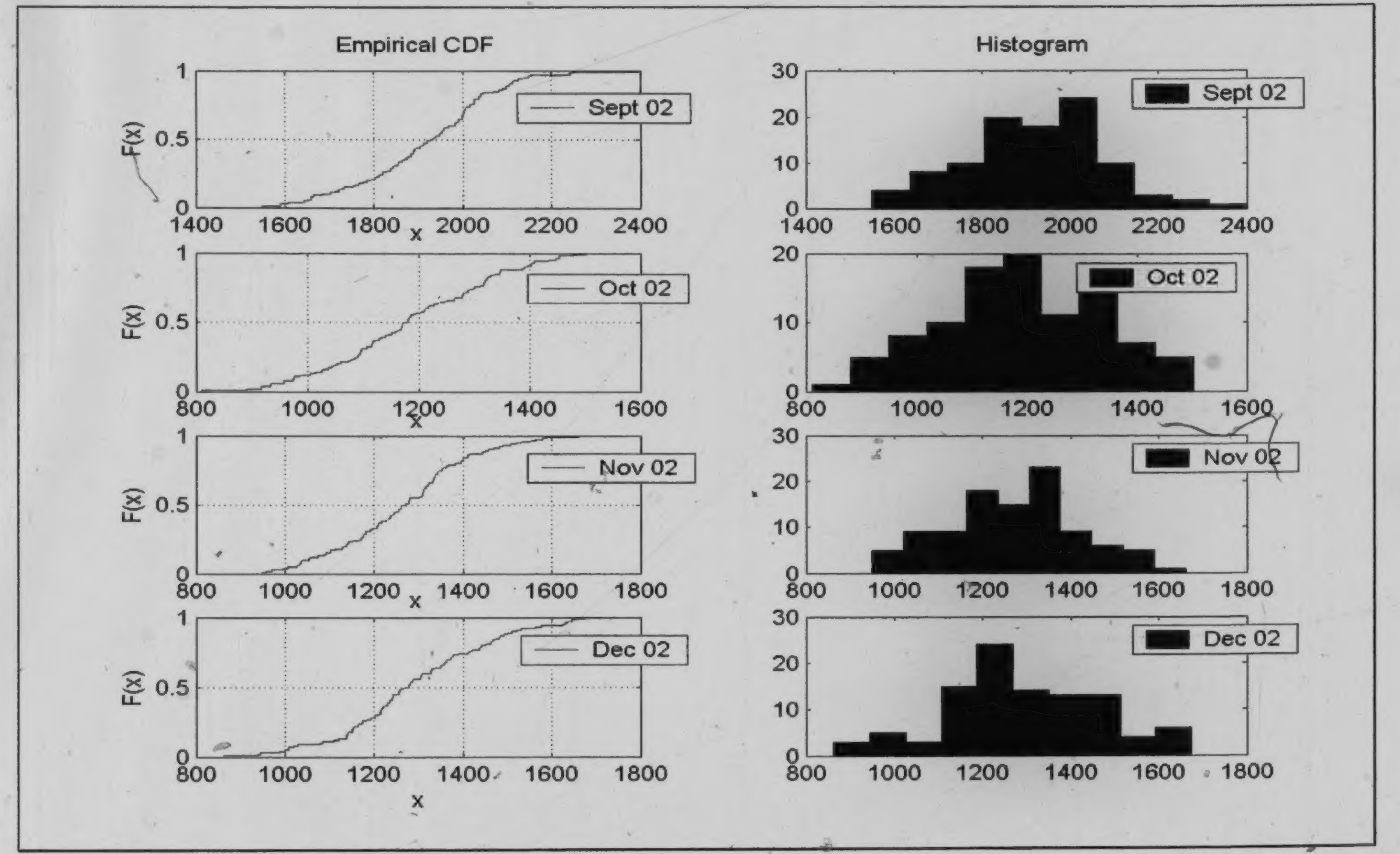


Figure 4-4: Empirical CDF and Histogram for Four Sample Time Points

Rank and Percent Analysis for Four Example Time Points

Sep-02	*Rank*	*Percent*	*Oct-02*	*Rank*	*Percent*	*Nov-02*	*Rank*	*Percent*	*Dec-02*	*Rank*	*Percent*
2399.1848	1	100.00%	1503.5371	1	100.00%	1660.7173	1	100.00%	1674.3687	1	100.00%
2248.5992	2	98.90%	1479.2859	2	98.90%	1588.7155	2	98.90%	1649.8183	2	98.90%
2244.1319	3	97.90%	1454.2146	3	97.90%	1580.0367	3	97.90%	1644.5688	3	97.90%
2154.3490	4	96.90%	1453.5372	4	96.90%	1563.3920	4	96.90%	1636.4744	4	96.90%
2148.1117	5	95.90%	1448.6491	5	95.90%	1561.2406	5	95.90%	1631.2147	5	95.90%
2145.5587	6	94.90%	1433.4321	6	94.90%	1520.4515	6	94.90%	1625.1280	6	94.90%
2124.9173	7	93.90%	1408.7689	7	93.90%	1508.7441	7	93.90%	1576.8919	7	93.90%
2115.8108	8	92.90%	1404.1067	8	92.90%	1489.5273	8	92.90%	1564.7098	8	92.90%
2112.8077	9	91.90%	1402.6425	9	91.90%	1483.2375	9	91.90%	1531.0378	9	91.90%
2108.6558	10	90.90%	1400.5430	10	90.90%	1470.2289	10	90.90%	1520.5421	10	90.90%
2104.1730	11	89.80%	1391.3967	11	89.80%	1457.1007	11	89.80%	1510.2469	11	89.80%
2097.0092	12	88.80%	1386.3231	12	88.80%	1448.9989	12	88.80%	1500.5309	12	88.80%
2096.6233	13	87.80%	1348.5399	13	87.80%	1444.9197	13	87.80%	1492.6218	13	87.80%
2089.2286	14	86.80%	1348.3856	14	86.80%	1415.6474	14	86.80%	1485.5570	14	86.80%
2081.8505	15	85.80%	1346.2904	15	85.80%	1411.9825	15	85.80%	1482.7274	15	85.80%
2069.1526	16	84.80%	1344.6560	16	84.80%	1409.9294	16	84.80%	1481.4798	16	84.80%
2047.7430	17	83.80%	1329.4892	17	83.80%	1401.1176	17	83.80%	1467.5210	17	83.80%
2042.6299	18	82.80%	1327.6801	18	82.80%	1400.2596	18	82.80%	1465.0617	18	82.80%
2042.5557	19	81.80%	1324.4815	19	81.80%	1396.7727	19	81.80%	1462.1615	19	81.80%
2037.7027	20	80.80%	1320.8941	20	80.80%	1395.3467	20	80.80%	1461.5086	20	80.80%
2036.4661	21	79.70%	1319.9508	21	79.70%	1383.0417	21	79.70%	1441.9163	21	79.70%
2027.3088	22	78.70%	1318.1305	22	78.70%	1368.5057	22	78.70%	1438.8181	22	78.70%
2027.0276	23	77.70%	1316.8911	23	77.70%	1361.9912	23	77.70%	1436.1051	23	77.70%
2026.6187	24	76.70%	1311.9695	24	76.70%	1360.9562	24	76.70%	1421.9344	24	76.70%
2023.1031	25	75.70%	1300.3182	25	75.70%	1351.4798	25	75.70%	1419.3281	25	75.70%
2016.5138	26	74.70%	1299.9751	26	74.70%	1346.0339	26	74.70%	1410.6672	26	74.70%
2011.9008	27	73.70%	1298.9094	27	73.70%	1345.8044	27	73.70%	1402.2406	27	73.70%
2006.7938	28	72.70%	1289.3740	28	72.70%	1343.1764	28	72.70%	1378.1006	28	72.70%
2003.1215	29	71.70%	1286.5543	29	71.70%	1339.6519	29	71.70%	1374.0263	29	71.70%
2003.0901	30	70.70%	1280.8910	30	70.70%	1337.8112	30	70.70%	1371.7709	30	70.70%
2002.3432	31	69.60%	1279.4482	31	69.60%	1336.9853	31	69.60%	1367.6193	31	69.60%
2000.4539	32	68.60%	1278.0378	32	68.60%	1335.5058	32	68.60%	1362.0561	32	68.60%
1998.5901	33	67.60%	1267.2253	33	67.60%	1333.3336	33	67.60%	1359.6705	33	67.60%
1998.0460	34	66.60%	1264.3622	34	66.60%	1331.0984	34	66.60%	1357.3176	34	66.60%
1997.7559	35	65.60%	1251.9779	35	65.60%	1324.9403	35	65.60%	1352.1584	35	65.60%
1991.8251	36	64.60%	1250.7405	36	64.60%	1323.5831	36	64.60%	1350.7988	36	64.60%
1989.3225	37	63.60%	1234.0127	37	63.60%	1322.5840	37	63.60%	1344.7058	37	63.60%
1988.7913	38	62.60%	1226.9065	38	62.60%	1316.6162	38	62.60%	1330.2230	38	62.60%
1984.0368	39	61.60%	1222.3721	39	61.60%	1314.4741	39	61.60%	1325.8418	39	61.60%
1982.0567	40	60.60%	1216.9710	40	60.60%	1312.1473	40	60.60%	1325.5866	40	60.60%
1968.4720	41	59.50%	1215.8908	41	59.50%	1311.8168	41	59.50%	1308.9349	41	59.50%
1960.0757	42	58.50%	1212.0240	42	58.50%	1311.0666	42	58.50%	1305.7167	42	58.50%
1956.1408	43	57.50%	1209.1947	43	57.50%	1308.4127	43	57.50%	1304.5460	43	57.50%
1954.4670	44	56.50%	1207.0856	44	56.50%	1307.7760	44	56.50%	1302.9287	44	56.50%
1951.7166	45	55.50%	1190.7247	45	55.50%	1301.9945	45	55.50%	1289.4410	45	55.50%
1949.7812	46	54.50%	1186.9082	46	54.50%	1282.0010	46	54.50%	1285.6289	46	54.50%
1948.2293	47	53.50%	1186.4310	47	53.50%	1281.2290	47	53.50%	1282.4250	47	53.50%
1938.4020	48	52.50%	1182.4775	48	52.50%	1279.1704	48	52.50%	1281.9975	48	52.50%
1936.1530	49	51.50%	1181.4186	49	51.50%	1277.8333	49	51.50%	1278.2191	49	51.50%
1932.6064	50	50.50%	1180.4886	50	50.50%	1271.8848	50	50.50%	1277.6947	50	50.50%

Table 4-4: Rank and Percent Analysis

Comparing these results with those from traditional forecasting, as provided in the previous section, one finds that Box Jenkins-based simulation provides much more information about future outcomes than traditional forecasting. Box Jenkins-based simulation also extends the forecasting functionalities into a new area. The following comparative table (Table 4-5) shows the differences:

Table 4-5: Comparison between Traditional Forecasting and Box Jenkins Simulation

Performance Criteria	**Traditional Forecasting**	**Box Jenkins Simulation**
Forecasts	Point estimate and range estimate	Mean, STD, Variance, Max, Min, etc.
Uncertainty Analysis	None	Risk and Uncertainty Analysis information, such as empirical CDF, histogram, percentile, probability.

4.1.3.2 Test on Dependent Simulation Input Modeling

The input modeling algorithms of Box Jenkins-based simulation is quite different from traditional statistical distribution-based input modeling. For statistical distribution-based input modeling, random variables are sampled from pre-fitted statistical distributions independently, without special considerations of the correlation between different sampling. As mentioned before, the wilful ignorance towards dependency in input modeling may cause a number of errors in the results. The proposed Box Jenkins-based simulation input modeling, therefore, does consider dependency by using a Box Jenkins time-series model as the input model.

To validate Box Jenkins-based simulation in providing dependent input modeling, the author performed a comparative test between traditional statistical distribution-based input modeling and Box Jenkins-based dependent input modeling. Graphical representations are used extremely to demonstrate the supposed dependency, which the traditional method does not consider but that the Box Jenkins based method successfully captures.

The following group of Scatter plot figures (Figure 4-5) show the correlations between variables with different lag (lag 1,2, and 3 for examples) for an original time-series, traditional sampled variables, and Box Jenkins sampled variables. The top three figures demonstrate the presence of a positive correlation in the original time-series. The simulation sampling based on the original data should, therefore, reflect the same correlation. The middle three figures are used to demonstrate a correlation among the simulated variates with different lag as generated using a traditional sampling technique. Obviously, however, there is not a strong correlation identified by these figures; therefore, it is clear that this technique fails in sampling dependent time-series data. The bottom three figures clearly show the same positive correlation among simulated variates with different lag as generated by Box Jenkins-based sampling technique as in the original time-series. Box Jenkins-based sampling is, therefore, capable of capturing the dependence.

Figure 4-6 shows a simple time-series plot for traditional method input modeling. Table 4-6 lists the tabulated descriptive statistics of simulation output analysis for traditional method of input modeling. If one compares Figure 4-3 with Figure 4-6 and Table 4-3 with Table 4-6, one will easily conclude that traditional distribution-based input modeling can neither generate dependent time-series data, nor produce an accurate estimation of future outcomes. The corresponding simulation analysis must, therefore, be incorrect and unreliable.

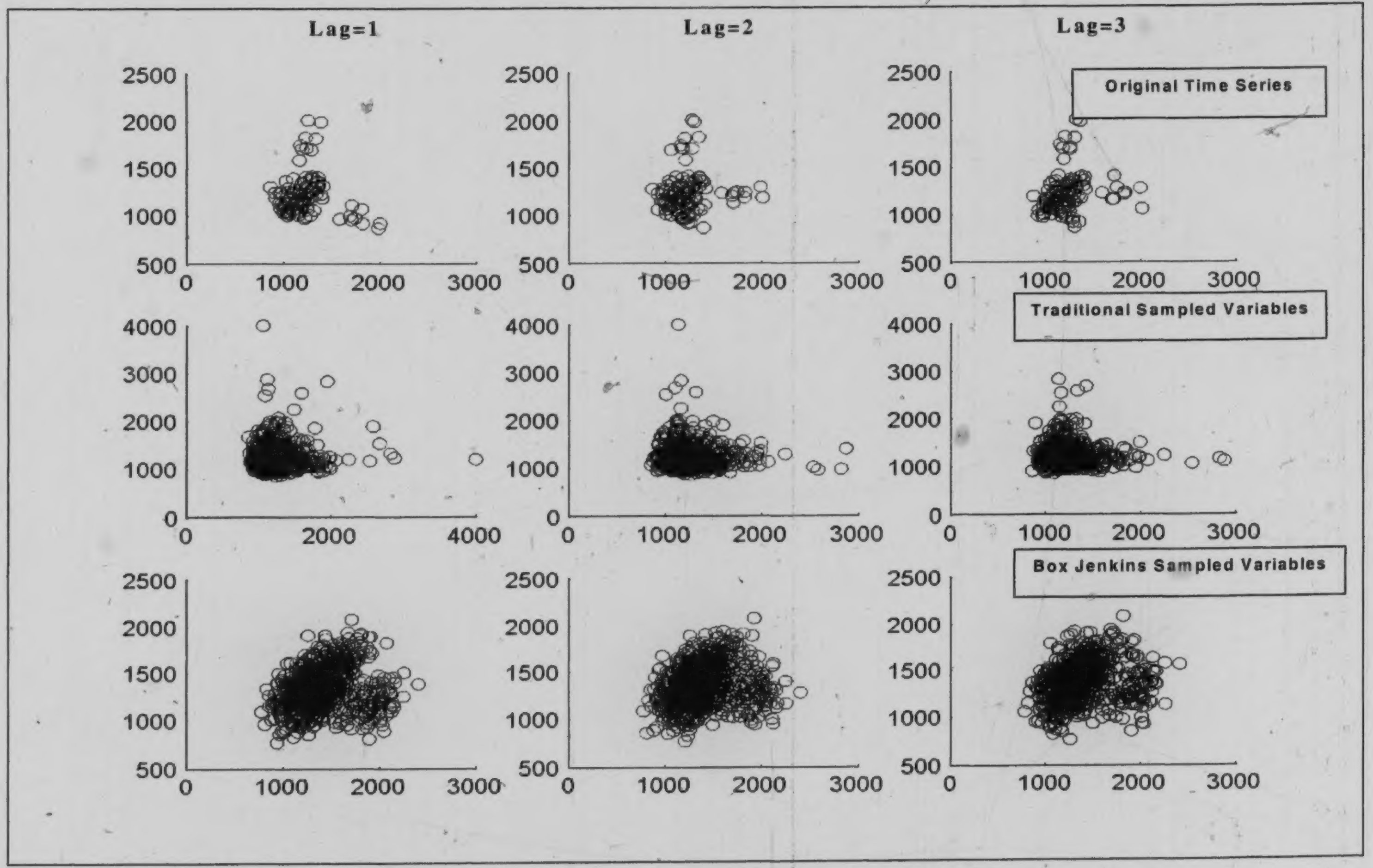


Figure 4-5: Scatterplot of Correlation between Observations with Different Lag (1, 2, and 3)

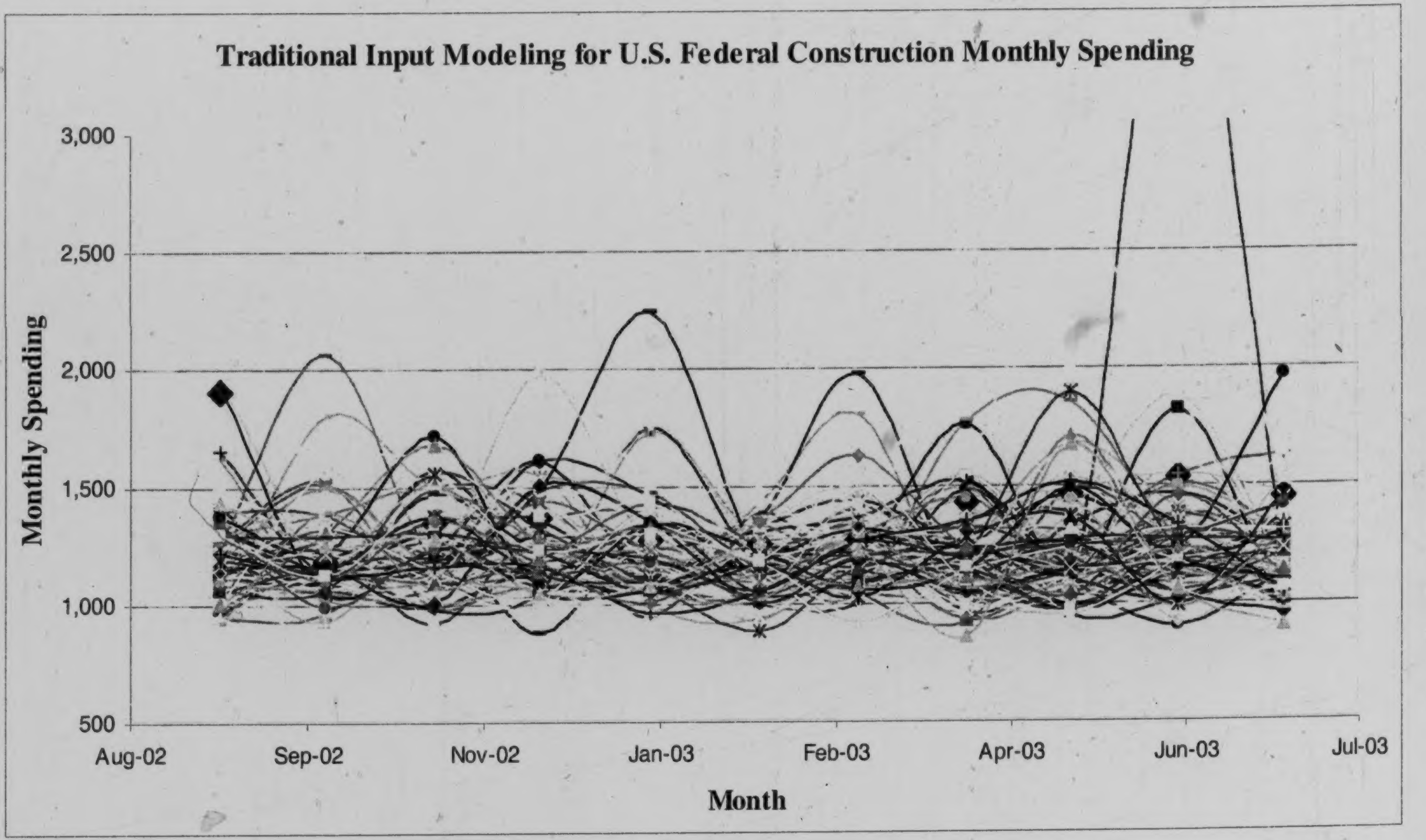


Figure 4-6: Time-series Plot for Traditional Input Modeling Method

Descriptive Statistical Analysis(Traditional Methodologies)

Sep-02		Oct-02		Nov-02		Dec-02	
Original	[illegible]	Original	[illegible]	Original	[illegible]	Original	[illegible]
Mean	1207.601	Mean	1228.454	Mean	1253.075	Mean	1234.630
Standard Error	16.35276332	Standard Error	24.67043744	Standard Error	19.91750311	Standard Error	17.97987422
Median	1196.625977	Median	1163.759338	Median	1193.232117	Median	1203.792358
Mode	#N/A	Mode	#N/A	Mode	#N/A	Mode	#N/A
Standard Deviation	163.5276332	Standard Deviation	246.7043744	Standard Deviation	199.1750311	Standard Deviation	179.7987422
Sample Variance	26741.28681	Sample Variance	60863.04834	Sample Variance	39670.69302	Sample Variance	32327.58768
Kurtosis	3.02521661	Kurtosis	19.91492554	Kurtosis	0.841162171	Kurtosis	2.175341675
Skewness	1.065673294	Skewness	3.636387573	Skewness	0.994403729	Skewness	1.10503487
Range	996.5276489	Range	1942.511353	Range	1049.356628	Range	1108.510681
Minimum	926.4338989	Minimum	922.2613525	Minimum	920.2954712	Minimum	875.046814
Maximum	1922.961548	Maximum	2864.772705	Maximum	1969.6521	Maximum	1983.557495
Sum	120760.0513	Sum	122845.4246	Sum	125307.4521	Sum	123463.009
Count	100	Count	100	Count	100	Count	100

Table 4-6: Descriptive Statistical Analysis (Traditional Methodologies)

4.2 Case Study

4.2.1 Introduction

The previous chapters have illustrated the details of Box Jenkins-based forecasting and simulation methodologies. Two validation tests have been conducted to verify the developed algorithms and methodologies. In order to demonstrate the strengths of Box Jenkins-based forecasting and simulation for dealing with the time related risk and uncertainty associated with engineering applications, a real project case was selected for study.

The City of Edmonton initialized the cost planning and budgeting of the South LRT Extension Project long before they started substantial site construction in 2003. The construction of the Section 1A project involves the tunnelling of underground railway and the installation of stations, utilities, and facilities. An Edmonton based engineering consulting company, the construction manager of this project, based on Edmonton, Alberta, identified the necessity of performing accurate medium-range cash flow prediction for this multi-disciplinary construction project. This type of prediction can serve not only as the basis for project planning and scheduling, but it can also provide a good reference for the project owner, the City of Edmonton, to well plan the capital budget for this project, which is collaborately funded by the City of Edmonton, Government of Alberta, and Government of Canada.

This consulting company has not managed this kind of project since the early stage LRT project was constructed in Edmonton. The available engineering data, therefore, dated back to more than 20 years ago. A proper inflation calculation must be applied to update these data for current cost conditions. Based on the information collected from tendering documents and from the construction site of the Section 1A project, This company created an annual cash flow prediction for the Section 1B project. The construction of this project will begin in March 2005.

This case study employed the proposed Box Jenkins forecasting and Box Jenkins-based Monte Carlo simulation methodologies to reattempt the prediction. The objective of this application is to verify the original forecasting with a comparison between original prediction results and those offered by Box Jenkins forecasting and simulation methodologies. The verification results were submitted to Stantec for comments and suggestions. The feedback can then serve as the basis for future research.

4.2.2 Problem Identification and Tentative Solution

The original cash flow prediction was based on a lump sum cost estimation of the Section 1B project. The whole project was then broken down into several sub-items, which are scheduled for construction in the following five years, according to project scheduling and the proposed capital budget.

To consider the timing characteristic of project cost along a five years time frame, the original cash flow analysis adopted a Future Worth Method (FWM) to convert present worth (i.e., worth in 2003) into real future worth. This method is sometimes called the Compound Inflation Method, insofar as the compound inflation rate is incorporated into the calculation of future value in order to include the fluctuation in money value along a time-axis.

The following equations show the compounded calculation of present value to future value:

$$F = \begin{cases} P(1+i)^N \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots (Equation\ 4-1) \\ P\prod_{n=1}^{N}(1+i_n) \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots (Equation\ 4-2) \end{cases}$$

Where P is the present value, F is the future value, i is the annual inflation rate, and N is the number of years under consideration. When i is equal annually, Equation 4-1 is used; otherwise, Equation 4-2 should be applied.

Three categories of cost inflation identified are supposed to vary over time. Three deterministic inflation rates were used to calculate the compounded rate. Future values for each year were obtained using the FWM. These values are summarized in Table 4-7.

Table 4-7: Inflation Rates for three Cost Items (Original Analysis)

Inflation Item	Inflation Rate
Construction Inflation	5%
Land Inflation	3%
Engineering Inflation	5%

The above calculations are based on an assumption that the inflation rate for each inflation item remains constant over time. This assumption and the corresponding practice within inflation calculation remains in popular use within the construction industry due to its simplicity and easy-of-use. In reality, inflation rates fluctuate considerably from year to year, so ignorance of this variation is not appropriate.

To solve this problem, a Box Jenkins-based forecasting and simulation model was built to estimate future inflation rates numerically and scientifically. A necessary risk analysis was performed to provide more information about future outcomes to assist in decision-making.

Inflation rate for a given time period can be calculated using the following formula (Boskers 2002):

$$\theta_k = \frac{CPI_{current}}{CPI_{previous}} - 1 \text{ -- (Equation 4-3)}$$

Where $CPI_{current}$ is the current period index, and $CPI_{previous}$ is the previous period index.

The estimation of inflation rates can, therefore, be conducted by forecasting the cost indexes for the corresponding inflation item within a five-year time frame. Three cost indexes are chosen to represent selected inflation items. The corresponding relationships are illustrated in Table 4-8.

Table 4-8: Selection of Cost Index for Corresponding Cost Inflation Item

Cost Index	Corresponding Cost Inflation Item
Construction Cost Index for 20 North American Cities Average (ENR)	Construction Cost Inflation
Engineering Cost Index for Total Engineering Cost, Alberta (Statistics Canada)	Engineering Cost Inflation
Landing Cost Index for Edmonton, Alberta (Statistics Canada)	Landing Cost Inflation

Due to the lack of specific and localized cost data, the above index data, which are collected annually by different organizations for construction industry, are chosen to represent the Province of Alberta or Edmonton capital region in terms of cost variation. However, these data reflect only the variation of corresponding cost item for general construction projects on average level. The author expected there would be a certain degree of errors by using these index data.

Three independent Box Jenkins forecasting and simulation analyses are performed based on the methodologies and algorithms introduced in previous chapters. The forecasting analysis is performed with the aid of a developed *Simphony* SPS template. The resulting ARIMA models, which include parameters used for simulation input modeling, performed 100 iterations of the simulation calculations. Those simulated variates generated by this simulation analysis, in the author's opinion, should reflect the real variation at a certain degree of accuracy.

4.2.3 Implementation and Results

The implementation of the case study was executed using a developed SPS template and MATLAB programs. Some output analysis was done using *Microsoft Excel*.

The following figures show the simulated random variates for three cost indexes along a time-line from 2004 to 2008. The total number of iterations is 100.

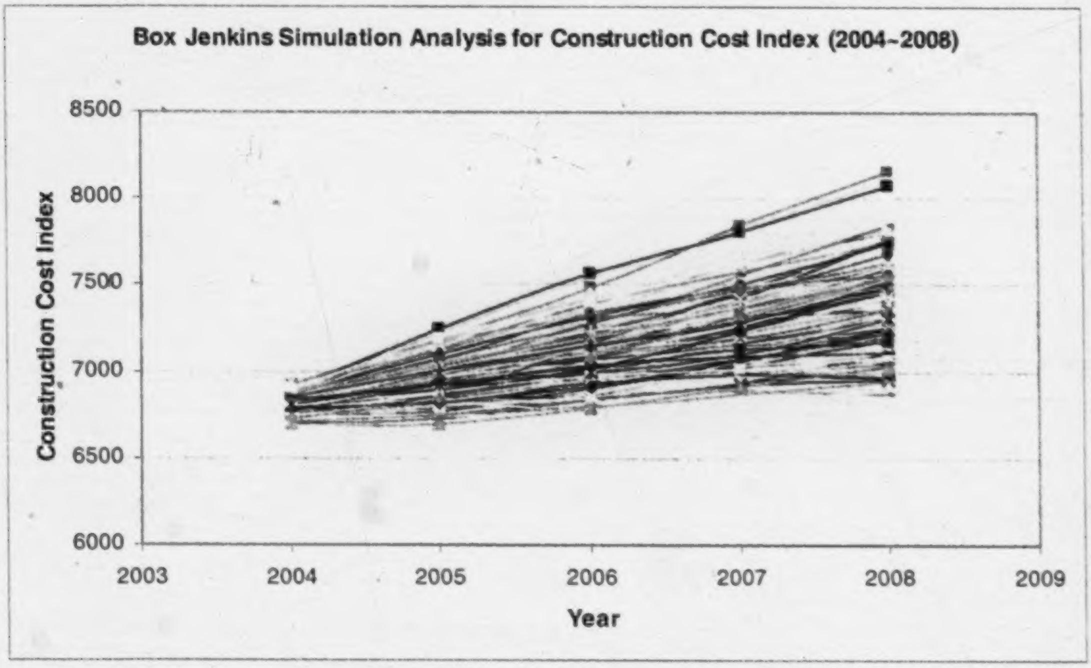


Figure 4-7: Simulated Random Variates for Construction Cost Index

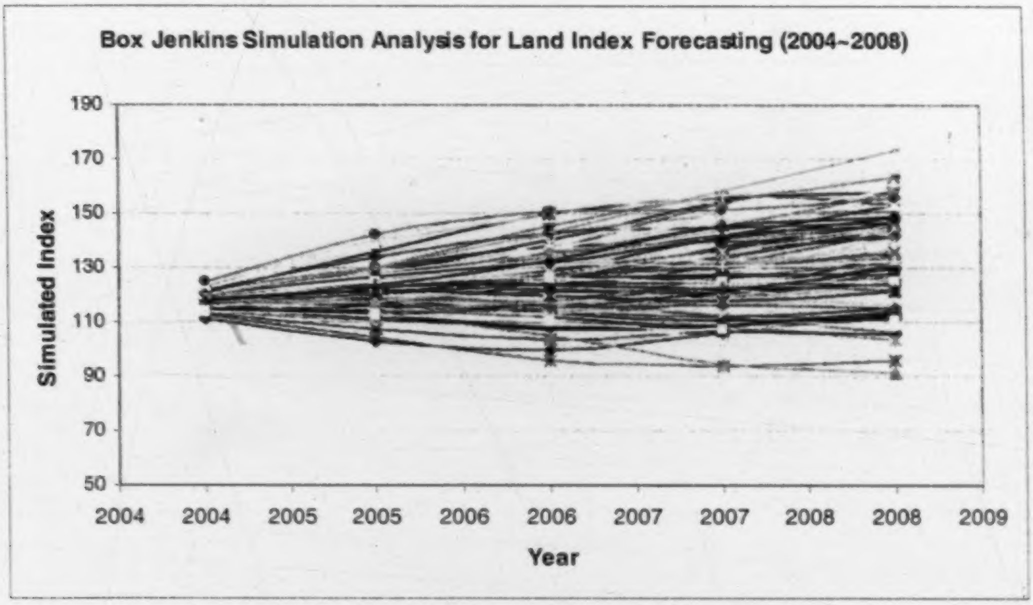


Figure 4-8: Simulated Random Variate for Landing Cost Index

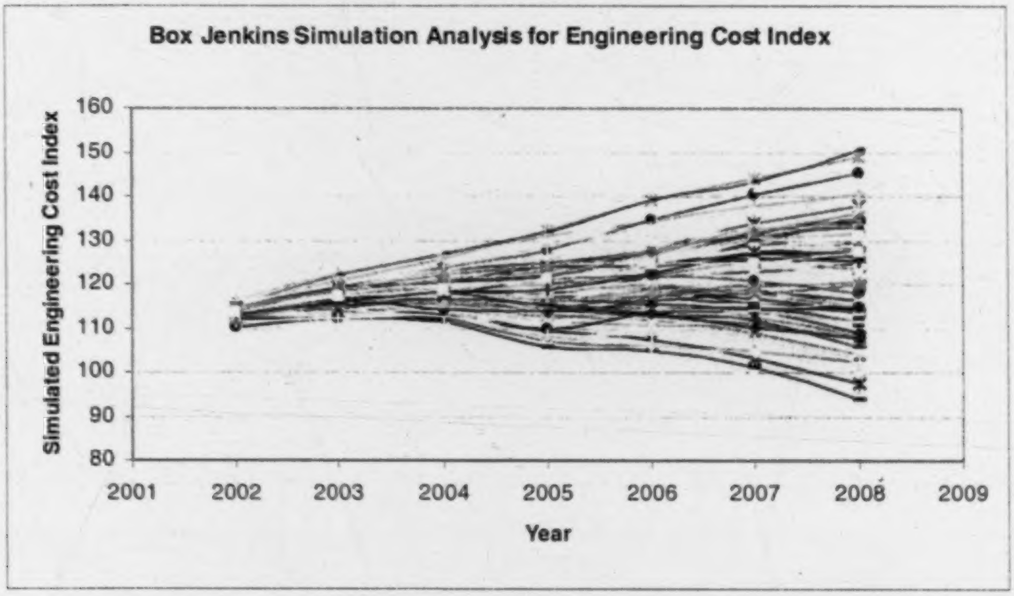


Figure 4-9: Simulated Random Variates for Engineering Cost Index

Based on the above-simulated random variates for the three cost indexes, the equations for the inflation rate calculation and the compounded future value method, a risk-based cash flow prediction was obtained.

The following tables (Table 4-9 and Table 4-10) tabulate the comparative analysis results between inflation figure estimations by the original forecasting and those by the risk-based Box Jenkins forecasting. Figure 4-10 through Figure 4-13 demonstrate the comparison graphically. For the purpose of confidentiality, the displayed annual cash flow data are scaled to hide the real data.

Table 4-9: Comparison of Inflation Rate Estimation

Comparison of Inflation Rate Estimation

Original Inflation Rate Estimation					
	2004	2005	2006	2007	2008
Construction Inflation	5.00%	5.00%	5.00%	5.00%	5.00%
Land Inflation	3.00%	3.00%	3.00%	3.00%	3.00%
Engineering Inflation	5.00%	5.00%	5.00%	5.00%	5.00%
Box Jenkins-based Inflation Rate Estimation					
50th Percentile					
Construction Inflation	1.85%	2.07%	2.17%	2.17%	2.16%
Land Inflation	3.06%	3.35%	2.64%	3.09%	3.23%
Engineering Inflation	1.72%	0.13%	1.76%	1.23%	-0.36%
60th Percentile					
Construction Inflation	2.05%	2.33%	2.38%	2.36%	2.37%
Land Inflation	3.59%	4.32%	3.51%	3.71%	3.58%
Engineering Inflation	1.71%	0.89%	2.11%	1.66%	0.20%
80th Percentile					
Construction Inflation	2.52%	2.93%	2.88%	2.80%	2.83%
Land Inflation	4.83%	6.56%	5.45%	5.05%	4.32%
Engineering Inflation	1.70%	2.61%	2.90%	2.60%	1.41%
90th Percentile					
Construction Inflation	2.88%	3.37%	3.25%	3.11%	3.17%
Land Inflation	5.75%	8.19%	6.82%	5.97%	4.82%
Engineering Inflation	1.68%	3.87%	3.46%	3.26%	2.24%
95th Percentile					
Construction Inflation	3.17%	3.73%	3.55%	3.37%	3.44%
Land Inflation	6.52%	9.52%	7.90%	6.68%	5.20%
Engineering Inflation	1.67%	4.90%	3.91%	3.79%	2.90%

Table 4-10: Comparison of Annual Cash Flow Estimation

Comparison of Cash Flow Estimation (Scaled)

Original Annual Cash Flow Estimation						
	2004	2005	2006	2007	2008	Total
Subjective Forecasts	$15,380	$79,214	$117,115	$201,264	$186,177	$599,149
Box Jenkins-based Annual Cash Flow Estimation						
Statistical Forecasts	2004	2005	2006	2007	2008	Total
50th Percentile	$14,993	$78,190	$108,717	$183,770	$167,004	$552,674
60th Percentile	$15,007	$78,720	$109,367	$185,072	$168,417	$556,584
80th Percentile	$15,053	$79,959	$110,936	$188,155	$171,755	$565,858
90th Percentile	$15,077	$80,873	$112,059	$190,406	$174,195	$572,610
95th Percentile	$15,098	$81,626	$112,983	$192,261	$176,204	$578,172

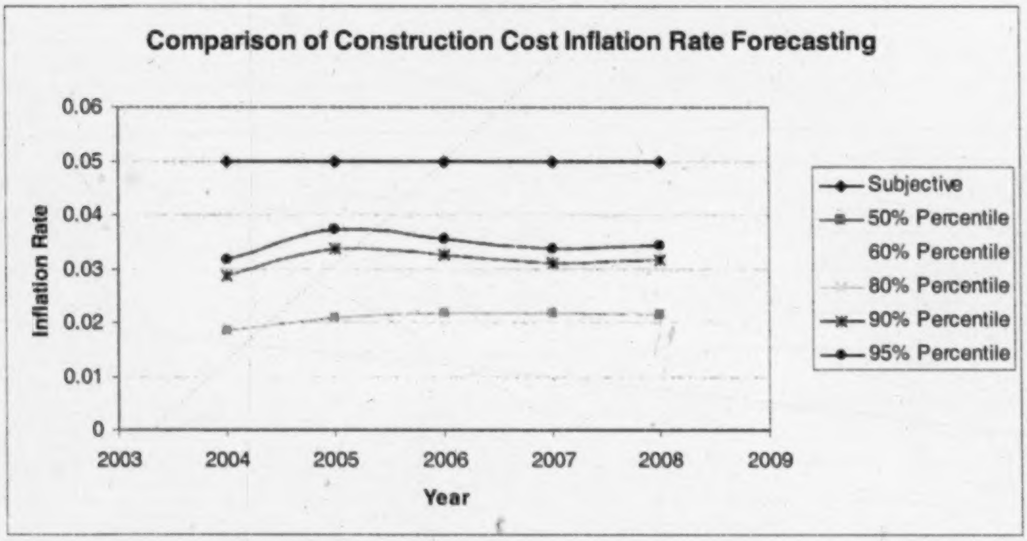


Figure 4-10: Comparison of Construction Cost Inflation Rate Forecasting

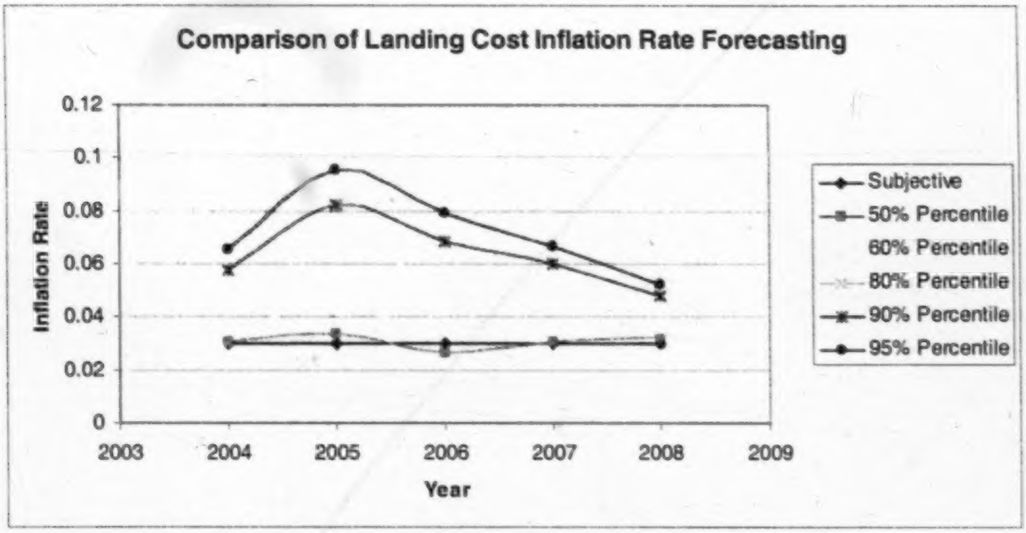


Figure 4-11: Comparison of Landing Cost Inflation Rate Forecasting

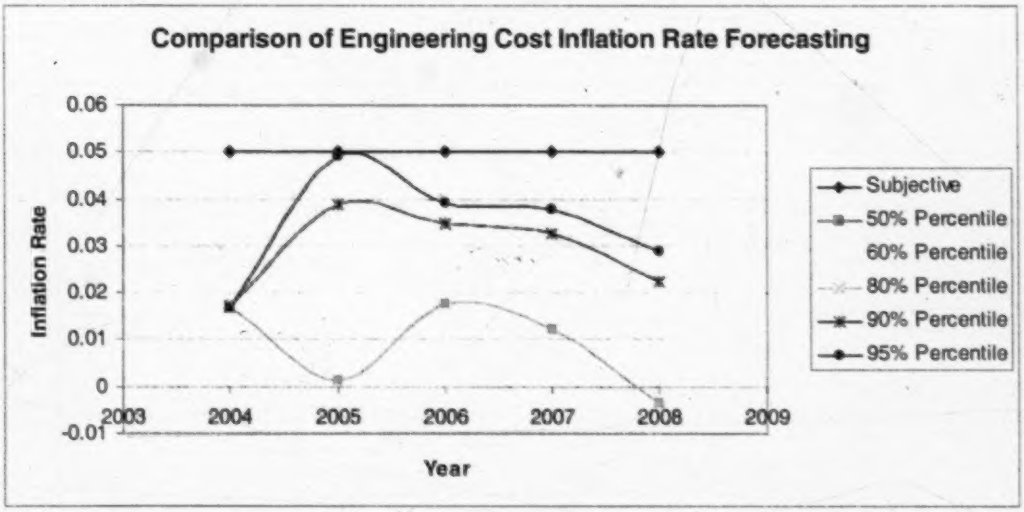


Figure 4-12: Comparison of Engineering Cost Inflation Rate Forecasting

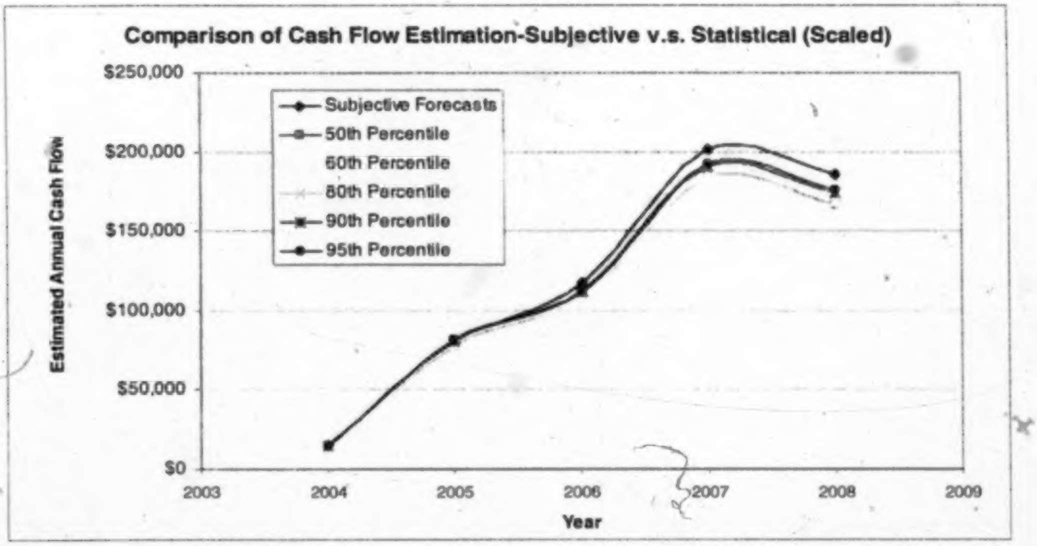


Figure 4-13: Comparison of Cash Flow Estimation - Subjective vs. Statistical (Scaled)

It is clear that Box Jenkins-based forecasting and simulation methods provide a more scientific estimation of the inflation rate variations and cash flow fluctuation than in the original analysis. It not only generated deterministic point forecasts, revealing that the cost inflation rates do not remain constant over time, but also provided risk analysis data such as the Maximum, Minimum, Mean, and Standard Deviation, in order to enrich the information already obtained. It provided a helpful risk analysis to the decision-maker involved in the capital planning task, in the form of the Arbitrary Quantile Points. With this information, project management professionals can perform risk-based analyses, such as scenario analysis. This scenario analysis will incorporate questions concerning the probability of meeting the projected cost requirements in 2004 if the total project budget for fiscal year 2004 is a particular amount and the minimum budget requirement if the decision-maker tolerates an 80% probability of project failure.

The above benefits arise from a unique risk analysis provided by Box Jenkins-based Monte Carlo simulation. Without simulation analysis, decision-makers will not have access to the same high level of information, which will render the decision-making process much more difficult and potentially inaccurate.

To get feedback from the industry about the case study, the analysis results were submitted to Stantec for comments and suggestions. The project manager of the South LRT extension project from this consulting company thoroughly reviewed the results and gave the following comments:

- Cash flow forecasting at the 90th percentile point is more useful than other percentile points for the purpose of project planning.
- The forecasted construction cost inflation rate is lower than the anticipated rate. The original one should be more realistic considering the specific market for underground subway construction in the Edmonton area, which fluctuate much more than the general construction market.
- The forecasted land cost inflation rate makes more sense than the original analysis. Accordingly an increase of the land cost inflation rate to 4% to 5% on average instead of the original 3% will be considered by Stantec.
- The forecasted engineering cost inflation rate is lower than the one that had been in use. This consulting company normally charges 5% for engineering cost inflation, which is more than the average level in the engineering industry. This charge rate is based on certain company business strategies.

In general, the project manager felt satisfied about the submitted analysis results and agreed that the objective of this case study, the verification of original cash flow forecasting was achieved.

4.3 Conclusion

This chapter discusses the research work on the verification of proposed automated Box Jenkins forecasting modeling tool and risk-based Box Jenkins forecasting model. The whole verification started from performing several validation tests on expected functionalities expected from the proposed modeling tool and the forecasting model, then an industry case study was utilized to test the applicability of proposed forecasting model in dealing with real capital planning problems.

The validation tests are performed in order to verify the superiority of Box Jenkins methodologies in forecasting functionality and simulation functionality. Comparative tests are conducted between Box Jenkins forecasting and other forecasting methodologies to show the advantages with regards to accuracy of Box Jenkins forecasting in providing short and medium-range forecasting. Box Jenkins-based simulation also proves to be more powerful in providing rich risk and uncertainty analysis information about the future outcomes than traditional forecasting method. As an academic contribution, Box Jenkins-based input modeling proves to be more adaptable in providing dependent sampling than traditional statistical distribution-based input modeling.

This case study applied Box Jenkins-based forecasting and simulation technologies to perform cash flow prediction and risk analysis for the City of Edmonton South LRT Section 1B Project. The analysis results were used to verify and improve the original cash flow prediction for project planning purposes. Compared to the original analysis from the local engineering consulting company, the new analysis derived the variation of project cost inflation from historical data, instead of from personal experience. It provided unique risk-based statistical information for project planning and scenario analysis, which enabled it to verify and improve upon the original analysis.

CHAPTER 5: CONCLUSIONS

5.1 Summary of Research

Generally, the capital planning of an infrastructure system is necessary due to shrinking project budget and an increased complexity of infrastructure projects. It is very difficult to be accurate in planning due to the high uncertainty of future outcomes and as a result of time and cost constraints. As one of the steps, which capital planning analysis should undertake, forecasting and estimating proves to be the most important, albeit the most expensive, step to acquire input information for further analysis. Based on a literature review, one of the reasons for which forecasting and estimating is expensive to perform, is that powerful, easy-to-use, and economical analysis tools are difficult to obtain despite the development of technologies.

This research was initialized to study the consequences of implementing advanced forecasting methodologies, such as Box Jenkins forecasting method, and to explore the possibility of simplifying these methodologies by automating originally manual modeling technique in the form of integrated computerized modeling. The research, subsequently determined the need for incorporating risk and uncertainty analysis into the existing method and to extend traditional forecasting into risk-based forecasting.

The development of Box Jenkins-based methodologies, which incorporates the automation of advanced forecasting modeling and the extension of current functionalities to include risk analysis, improves current analysis tools by offsetting the limitations of single forecasting particularly in accuracy and adaptability. It also employs simulation technologies to provide terms of risk-based statistical analysis for future estimation. The methodology has also proven to be, especially easy to use with the assistance of the developed automatic and integrated modeling tool and other proposed computer programs. This practical aspect is an important concern for average users in the construction industry.

All the theories and algorithms associated with this research are coded and tested using *MATLAB* 13.0. The related computer codes were developed in order to facilitate the implementation of these methodologies by average users. The automation of Box Jenkins forecasting methodologies is realized by the application of a SPS template developed in *Simphony*. This template also becomes a part of the implementation tools for Box Jenkins-based Monte Carlo simulation.

The necessary validation tests for the proposed methodologies were performed using actual engineering data. A comparative analysis between traditional methodologies and Box Jenkins-based methodologies strongly demonstrated the advantage of Box Jenkins forecasting and simulation in providing accurate forecasts and unique risk analysis. Comparative testing also proved that Box Jenkins-based simulation input modeling provides a feasible solution for dependent input modeling, which traditional statistical distribution-based input modeling is unable to provide.

The South LRT Extension Section 1B project annual cash flow prediction was studied to demonstrate the adaptability of Box Jenkins forecasting and simulation technologies in a real construction management case. Based on an analysis of the results, one can clearly observe the difference between the original analysis and the new analysis in terms of the increased availability of accurate estimation and risk analysis information. The feedback from the industry, regarding the analysis results, verified that these technologies provided a more scientific and risk-based prediction. The suggestions also emphasized the importance of using more customized and localized historical data for analyses specifically designed for a particular geographical area in a particular period of time.

In general, Box Jenkins-based forecasting and simulation technologies provide more advanced and generic estimation tools for capital planning problems.

5.2 Research Contributions

Contributions to the state of the art in capital system planning

The currently available methodologies for capital system planning were discussed in Chapter 2; the several limitations associated with these methodologies were listed. This research enabled the objectives that initiated this project to be achieved successfully. The applicable risk-based forecasting techniques, using Box Jenkins methodologies, are provided for short and medium-range capital planning tasks. These methodologies also prove effective for long-range capital planning tasks, particularly as compared to the available methodologies.

Two specific contributions to the state of the arts in capital system planning are identified:

- The simplification and automation of complex and advanced forecasting methodology applications for solving engineering problems for non-professionals.

This contribution resulted in a significant reduction of overall time and cost spent in the application of Box Jenkins-based methodologies as a result of using the developed *Simphony* SPS Template.

- The combination of advanced forecasting methodologies with simulation technologies to incorporate risk and uncertainty analysis into traditional forecasting analysis.

This research provides a good starting point for continued research and development of the combination of forecasting method and simulation technologies.

Contributions to the academic research on construction simulation technologies

- This research work developed the input modeling algorithms for univariate random process modeling with time series data

A complete set of input modeling algorithms, including all the methodologies, modeling procedures, and implementation tools were developed for the application of Box Jenkins-based Monte Carlo simulation input modeling, simulation execution, and output analysis.

- The research developed dependent input modeling algorithms for univariate time-series data.

This research provided an applicable solution for proposed dependent random sampling, using Box Jenkins forecasting models. Compared with traditional statistical distribution-based random sampling, this solution has the ability to adapt to dependent time-series data.

5.3 Limitations of Research and Related Future Work

This research was initialized under the topic of *Univariate Random Process Input Modeling*, and focused on capital planning using Box Jenkins methodologies. A complete set of input modeling algorithms for univariate time-series data has already been established and the related software for its implementation has been developed. In a real situation, however, there are always several time-related factors. These factors vary over time and interact among themselves, making the risk and uncertainty associated with the correlation more complex than situations that involve only a single factor. Accordingly, this research has the limitation of focusing on unvariate input modeling without any consideration of the multivariate cases that might present unforeseen influencing factors or of assuming a single factor will dominate an entire situation in a particular case.

Another limitation exists with the output analysis that follows input modeling and simulation calculations in a complete simulation analysis. In this research, the author employed the output analysis methods established by AbouRizk and Halpin (1990) for traditional statistical distribution-based input modeling methodologies. These methods are not tested for the adaptability of them on Box Jenkins-based input modeling methodologies. The results, therefore, are immature and vulnerable to in-depth study. In the view of the time constraints, this research only focuses on the study of input modeling and the development of related algorithms and methodologies. The established output analysis methods were taken for granted. It is still necessary to perform a detailed study on output analysis methodologies for Box Jenkins-based Monte Carlo simulation experiments.

A third limitation exists with the implementation of Box Jenkins-based Monte Carlo Simulation. Although the input modeling has been automated using the developed *Simphony* SPS template, the simulation execution and output analysis still require users to operate several computer programs developed in the *MATLAB* 13.0 environment.

These limitations indicate the direction that future research may take, which are summarized as follows:

- Multi-variate input modeling algorithms for time-series data must be studied and associated implementation methods must likewise be developed. A correlation analysis between different time-series and the impact of correlation on simulation output analysis may prove to be a helpful topic for research.
- The output analysis for Box Jenkins-based Monte Carlo simulation should be studied in detail and associated methodologies should be established.
- The automation of a complete Box Jenkins-based Monte Carlo simulation should be realized along with the development of an integrated computer simulation system.

REFERENCES AND BIBLIOGRAPHY

AbouRizk, Simaan M. Halpin, Daniel W. Lutz, James D. (1992), "State of the Art in Construction Simulation", *Proceedings of the 1992 Winter Simulation Conference,* Arlington, Virginia, United States, 1271-1277.

AbouRizk, Simaan M. and Halpin, Danniel W. (1990), "Probabilistic Simulation Studies for Repetitive Construction Processes", *Journal of Construction Engineering and Management,* 116(4), 575-594.

Bashford, Howard H. et al. (2003), "Implications of Even Flow Production Methodology for U.S. Housing Industry", *Journal of Construction Engineering and Management*, 129(3), 330-337.

Biller, Bahar and Nelson, Barry L. (2003), "Modeling and Generating Multivariate Time-Series Input Processes Using a Vector Autoregressive Technique", *ACM Transaction on Modeling and Computer Simulation,* 13(3), 211-237.

Boskers, Nathan David (2002). "Modeling Uncertainty in Capital Construction Projects with Lengthy Implementation Time Frames". MSc Thesis, *Construction Engineering and Management Group,* U of Alberta, Edmonton, Canada.

Cario, Marne C. (1996), "Modeling and Simulation Time Series Input Processes with ARTAFACTS and ARTAGEN", *Proceedings of the 1996 Winter Simulation Conference,* Coronado, CA, USA, 207-213.

Enright, Michael P. and Frangopol, Dan M. (1998), "Service-Life Prediction of Deteriorating Concrete Bridges", *Journal of Structural Engineering,* 124(3), 309-317.

Harding, Benjamin L. and Walski, Thomas M. (1996), "Long Time-Series Simulation of Water Quality in Distribution Systems", *Hydrosphere, Resource Consultants, Inc., Wilkers University*

Ingalls, Ricki G. (2001), "Introduction to Simulation", *Proceedings of the 2001 Winter Simulation Conference,* Arlington, VA, USA, 7-16.

Leemis, Lawrence (2001), "Input Modeling Techniques for Discrete-Event Simulations", *Proceedings of the 2001 Winter Simulation Conference,* Arlington, VA, USA, 62-73.

Li, Jenny X. and Winker, Peter (2000), "Time Series Simulation with Quasi Monte Carlo Methods", *Department of Mathematics and Economics, PSU, Department of Economics, U of Mannheim.*

Li, Y. et al. (2003), "Statistical forecasting of soil dryness index in the southwest of Western Australia", *Forest Ecology and Management, 183(1-3), 147-157*

Nelson, Barry L. and Yamnitsky, Michael (1998), "Input Modeling Tools for Complex Problems", *Proceedings of the 1998 Winter Simulation Conference,* Washington DC, USA, 105-112.

Schmeiser, Bruce (1999), "Advanced Input Modeling for Simulation Experimentation", *Proceedings of the 1999 Winter Simulation Conference,* Phoenix, AZ, USA, 110-115.

Sun, Hongbing and Koch, Manfred (2001), "Case Study: Analysis and Forecasting of Salinity in Apalachicola Bay, Florida, Using Box Jenkins ARIMA Models", *Journal of Hydraulic Engineering, 1*27(9), 718-727.

Tang, Y. F. Lam William H. K. and Pan L. P. (2003), "Comparison of Four Modeling Techniques for Short-Term AADT Forecasting in Hong Kong", *Journal of Transportation Engineering*, 129(3), 271-277.

Wilson, James R. (1997), "Modeling Dependencies in Stochastic Simulation Inputs", *Proceedings of the 1997 Winter Simulation Conference*, Coronado, CA, USA, 47-52.

AC/UNU Millennium Project (1994), "Statistical Modeling: From Time Series to Simulation", *The Futures Group*.

Canada, John R., et al (1996). *Capital Investment Analysis for Engineering and Management*, Prentice Hall, Inc. Upper Saddle River, New Jersey.

Erdman, Donald (2002), "Simulation Models for Business Planning and Economic Forecasting", *SAS/ETS User's Guide*, SAS Institute Inc., Cary, NC.

Gaynor, Patricia E. and Kirkpatrick, Rickey C. (1994), *Introduction to Time Series Modeling and Forecasting in Business and Economics*, McGraw-Hill, Inc.
MATLAB 13.0 Documentations (2003), *the MathWorks, Inc.*

McGraw-Hill Inc. Construction, *ENR- Construction Cost Index*, http://www.enr.com/.
NSERC/Alberta Construction Industry Research Chair, *Guide to The Development of Special Purpose Simulation Template with Simphony*, U of Alberta, Edmonton.

NSERC/Alberta Construction Industry Research Chair, *User's Guide to Simphony*, U of Alberta, Edmonton.

Pankratz, Alan (1983). *Forecasting With Univariate Box-Jenkins Models*. John Wiley & Sons, Inc.

Statistics Canada, *Engineering Cost Index for Alberta and Landing Cost Index for Edmonton*, http://www.statcan.ca/start.html.

U.S. Census Bureau. *Public Construction Spending Statistics*, http://www.census.gov

APPENDIX A: Validation Study for Box Jenkins Forecasting Methodologies

In order to validate the usability and adaptability of Box Jenkins forecasting methodologies, four basic forecasting methods are selected for a comparative analysis. The detailed modeling procedures of each method are attached below for reference:

1. Moving Average

Based on the introduction about moving average method from *Time-series Modeling and Forecasting in Business and Economics* (Gaynor and Kirkpatrick 1994), this method is a basic methodology to forecast the future outcomes through the averaging of most recent values. The methodology, therefore, is not difficult to be understood and be implemented. However, the selection of how many most recent periods (numbered by *k*) to utilized in computing the moving average is determined by the users. Normally, this number is determined by trial and error method in order to choose the *k*-period moving average forecast that generates the smallest error.

In this test, a collection of 12 data points from a testing data set was selected for the identification of *k* value. Tentative *k* values from three to thirteen are attempted and the resulting forecasts for these 12 periods through moving average method are compared. The following Table A-1 demonstrates the comparison.

Based on the content of Table A-1, it is obvious that moving average forecasts for the selected 12 data points using *k* value as 12 produce the most accurate estimation and therefore k value as 12 was selected to be used in the following formal forecasting.

The final forecasts on the succeeding 11 periods (September 2002 to July 2003) through moving average method were provided in Chapter 3. The following time-series plot (Figure A-1) shows the original values against the forecasts along the time axis.

Table A-1: Table for the selection of *K* value for moving average method

Period	Original	K	APE	K	APE	K	APE	K	APE	K	APE	K	APE	K	APE	K	APE
		3		4		5		6		7		8		12		13	
Sep-01	1814.2	1280.6	29.4%	1272.2	29.9%	1254.3	30.9%	1249.8	31.1%	1231.4	32.1%	1223.9	32.5%	1237.3	31.8%	1241.4	31.6%
Oct-01	1073	1273.4	18.7%	1278.5	19.1%	1268.6	18.2%	1253.6	16.8%	1247.2	16.2%	1230.5	14.7%	1199.4	11.8%	1237.6	15.3%
Nov-01	1188	1475.8	24.2%	1408.1	18.5%	1384.9	16.6%	1359.4	14.4%	1333.3	12.2%	1318.0	10.9%	1264.0	6.4%	1246.7	4.9%
Dec-01	1213	1454.3	19.9%	1408.8	16.1%	1362.4	12.3%	1348.1	11.1%	1329.9	9.6%	1309.6	8.0%	1258.4	3.7%	1253.0	3.3%
Jan-02	1144	1238.4	8.3%	1371.0	19.8%	1355.5	18.5%	1327.2	16.0%	1321.0	15.5%	1309.2	14.4%	1260.4	10.2%	1252.1	9.5%
Feb-02	1264	1213.1	4.0%	1211.2	4.2%	1328.7	5.1%	1325.9	4.9%	1307.0	3.4%	1304.9	3.2%	1264.0	0.0%	1256.3	0.6%
Mar-02	1313	1190.0	9.4%	1189.1	9.4%	1189.3	9.4%	1293.0	1.5%	1297.0	1.2%	1284.6	2.2%	1266.3	3.6%	1254.5	4.5%
Apr-02	1283	1199.3	6.5%	1202.5	6.3%	1199.7	6.5%	1195.8	6.8%	1284.7	0.1%	1290.1	0.6%	1269.5	1.1%	1265.4	1.4%
May-02	1373	1258.8	8.3%	1230.9	10.4%	1226.7	10.7%	1219.6	11.2%	1211.4	11.8%	1287.4	6.2%	1280.6	6.7%	1272.8	7.3%
Jun-02	1353	1284.9	5.0%	1272.7	5.9%	1246.1	7.9%	1239.4	8.4%	1230.9	9.0%	1220.7	9.8%	1284.5	5.1%	1281.1	5.3%
Jul-02	1364	1313.6	3.7%	1310.4	3.9%	1295.8	5.0%	1269.4	6.9%	1260.1	7.6%	1249.8	8.4%	1290.8	5.4%	1291.4	5.3%
Aug-02	1404	1346.5	4.1%	1329.9	5.3%	1323.6	5.7%	1309.2	6.7%	1284.3	8.5%	1274.1	9.3%	1304.1	7.1%	1296.1	7.7%
MAPE			11.8%		12.4%		12.2%		11.3%		10.6%		10.0%		7.7%		8.1%

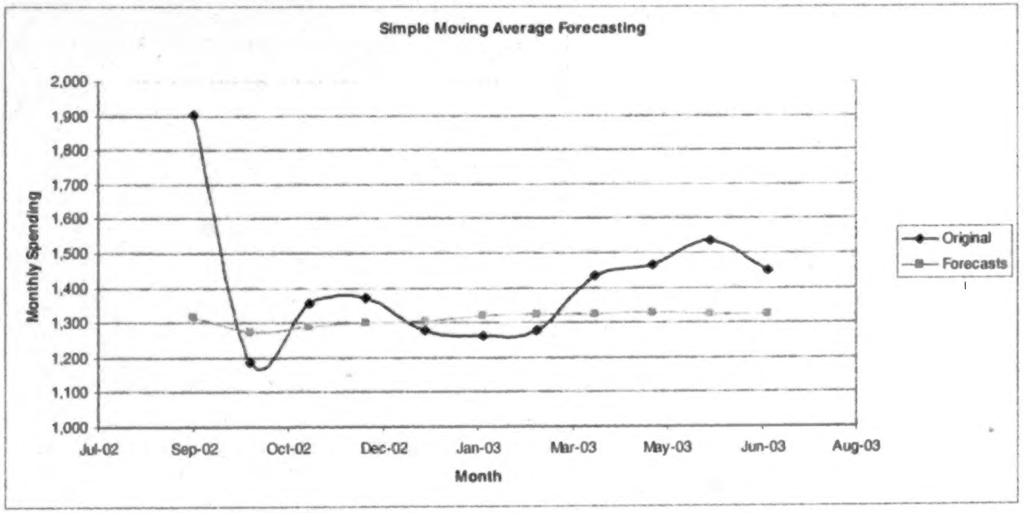


Figure A-1: Moving Average Forecasts against Original Data

2. Regression

The essence of regression method is to identify independent factors for dependent factor, thus a linear or non-linear regression function can be established between independent factors and dependent factor.

However, in this case, the author found that it is difficult to identify what kinds of independent factors can cause the variation of federal monthly non-residential spending based on the currently available information. It is also impossible to get the historical data for these factors even if some factors are artificially figured out because these factors may not have numerical forms.

The author, therefore, chose the simple trend forecasts, which means one can use time as an independent factor in the regression relationship function and the federal monthly non-residential spending as the left hand side dependent factor. In this way, a simple linear regression function can be established:

$$Y_t = \alpha + \beta t + \varepsilon_t \text{ -- (EquationA-1)}$$

Where Y_t = actual values for the time series variable;

t = time (1,2,3,…,n);

α , β = coefficients for the regression function; and

ε_t = random error in time period t.

The forecasts values for Y_y are determined by estimating of the regression coefficients for α and β and then by solving the equation for each value of t.

The coefficient estimation results are tabulated as follows:

Table A-2: Summary of Trend Regression Analysis

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.039146896							
R Square	0.001532479							
Adjusted R Square	-0.007226008							
Standard Error	205.3834115							
Observations	116							
ANOVA								
	df	*SS*	*MS*	*F*	*ignificance F*			
Regression	1	7380.679	7380.679	0.174971	0.676518			
Residual	114	4808787	42182.35					
Total	115	4816168						
	Coefficients	*tandard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
Intercept	1211.182564	38.38667	31.55217	3.68E-58	1135.139	1287.226	1135.139	1287.226
X Variable 1	0.238214354	0.569489	0.418295	0.676518	-0.889939	1.366368	-0.889939	1.366368

From the above table, one can observe the bad performance of regression method in fitting the original data. The *R* square, which is used as a major performance index for fitting quality, is just about 0.1%. This value is unacceptable based on forecasting requirements. The following figure (Figure A-2) can also demonstrate the bad fitting performance. The resulting forecasts have been provided in the related discussion in Chapter 3.

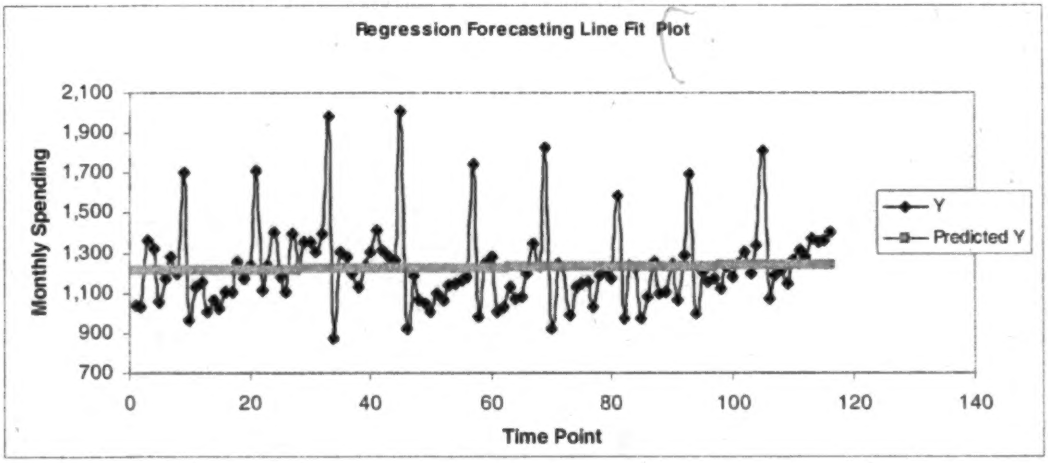


Figure A-2 Data Fitting for Trend Regression Method

3. Exponential Smoothing

There is no big difference between basic exponential smoothing and moving average in terms of foundational mathematical calculation. As an improvement on methodology, exponential smoothing gives different weights on different recent values in considering the different degree of effects of recent values on future outcomes when performing the averaging operations.

The basic exponential smoothing model one use is as follows:

$S_t = \alpha' x_t + (1-\alpha')S_{t-1} \quad (0 \le \alpha' \le 1)$ -- (Equation A-2)

Or

(Forecast for period t+1, made in period t)= α' (Actual data point in period t) + (1+ α')(Forecast for period t, made in period t-1)------------------------------(Equation A-3)

Based on the rule of thumb for the selection of α' (Smoothing Constant) or (1- α') (Damping Factor), α' should lie between 0.01 and 0.3. In this case, the author chooses 1- $\alpha' = 0.3$.

Exponential smoothing has proven incapable of providing the accurate forecasts for the testing time-series data. The reason, based on the study on the forecasting, is that during the weighted averaging calculation process, there is no further real observation, which can be obtained to modify the forecast and offset the error, therefore one can only assume a current forecast is the approximation of a real observation. Ultimately the forecasting error is accumulated and the resulting forecasts, accordingly, are far way from the real values.

The following figure (Figure A-3) demonstrates the original data against the forecasts through exponential smoothing method.

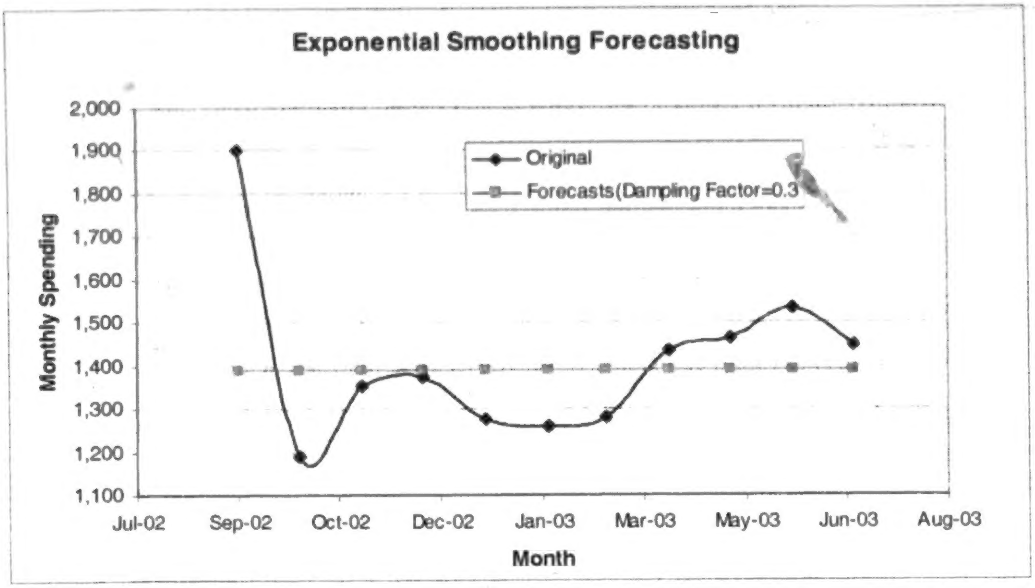


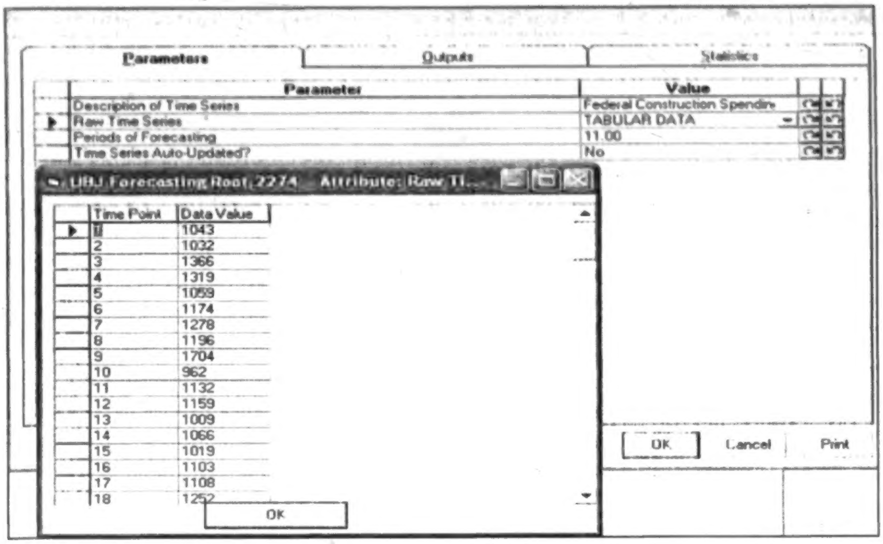
Figure A-3: Forecasts from Exponential Smoothing Method

4. Box Jenkins

The execution of Box Jenkins forecasting methodologies is assisted using the developed SPS *Simphony* template. The detailed instruction of this template can be referred to from Appendix D.

The following series of figures illustrate the automatic modeling process progressively. These pictures come from the screenshot of the executed simulation model using the developed SPS template under *Simphony* environment.

1) Data Input

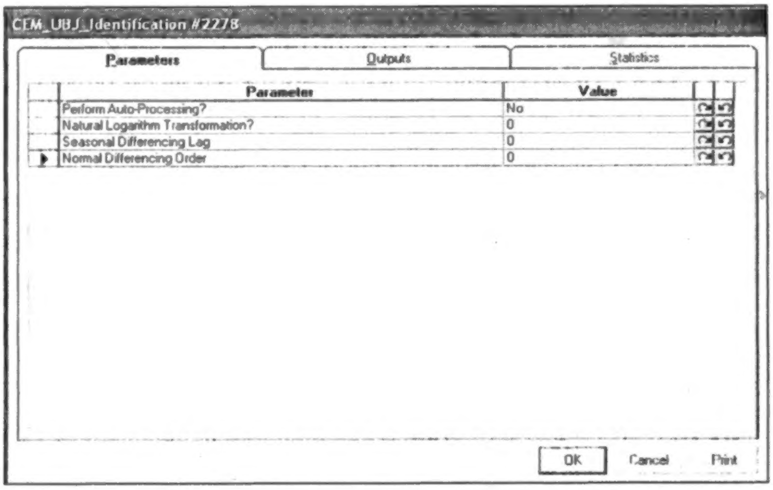

Parameters
Outputs
Statistics
UBJ Forecasting Root, 2274 Attribute: Raw Ti...
OK
OK
Cancel
Print

Parameter	Value
Description of Time Series	Federal Construction Spendin
Raw Time Series	TABULAR DATA
Periods of Forecasting	11.00
Time Series Auto-Updated?	No

Time Point	Data Value
1	1043
2	1032
3	1366
4	1319
5	1059
6	1174
7	1278
8	1196
9	1704
10	962
11	1132
12	1159
13	1009
14	1066
15	1019
16	1103
17	1108
18	1252

Figure A-4: Screenshot of Data Input Window

2) Modeling Initialization


CEM_UBJ_Identification #2278
Parameters
Outputs
Statistics
OK
Cancel
Print

Parameter	Value
Perform Auto-Processing?	No
Natural Logarithm Transformation?	0
Seasonal Differencing Lag	0
Normal Differencing Order	0

Figure A-5: Screenshot of Modeling Initialization Window

3) Judgement of Stationary

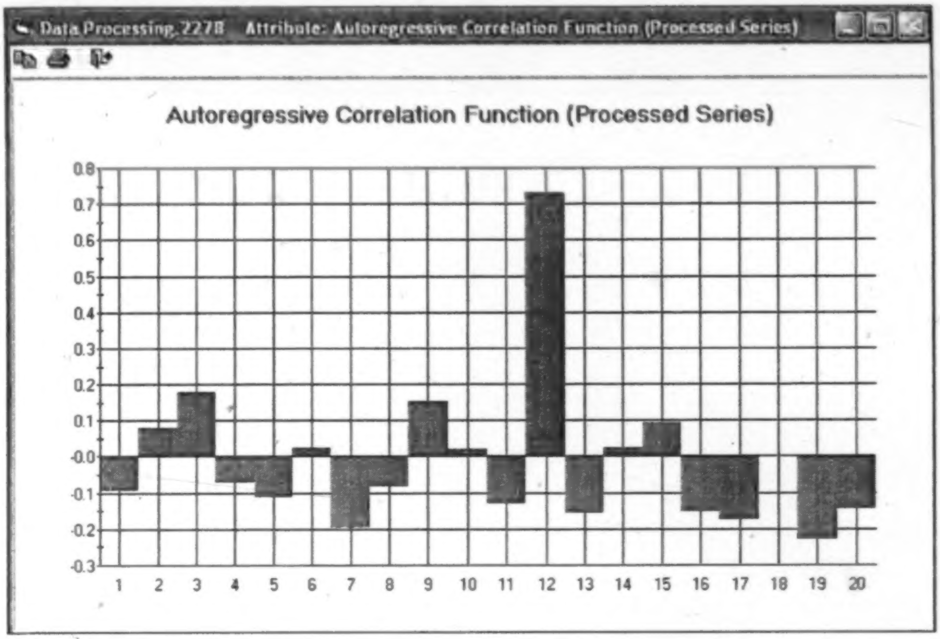

Data Processing.2278 Attribute: Autoregressive Correlation Function (Processed Series)
Autoregressive Correlation Function (Processed Series)
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
-0.0
-0.1
-0.2
-0.3
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20

Figure A-6: Screenshot of ACF Plot Window

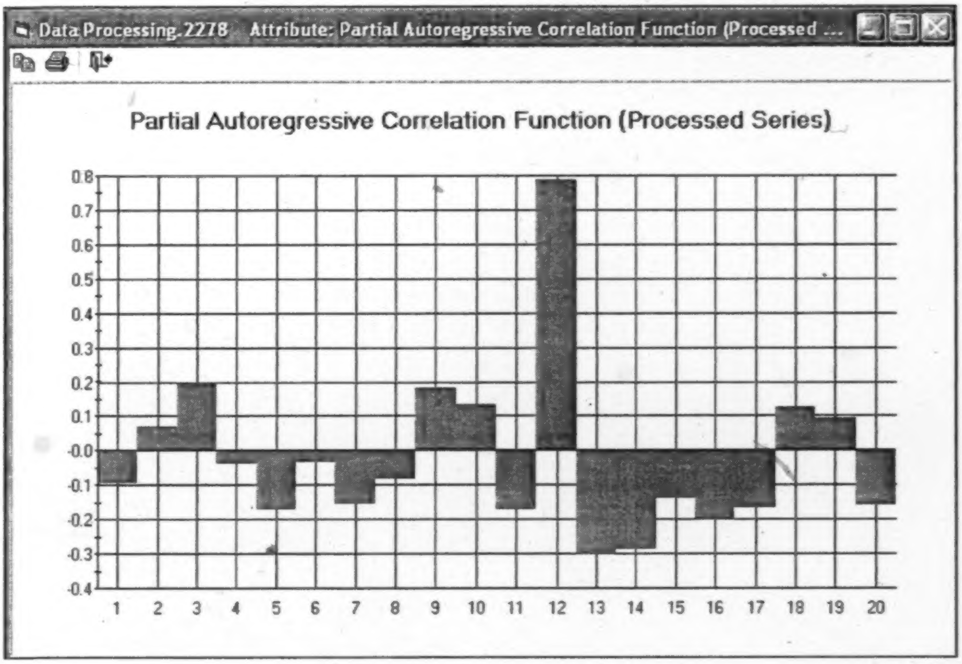
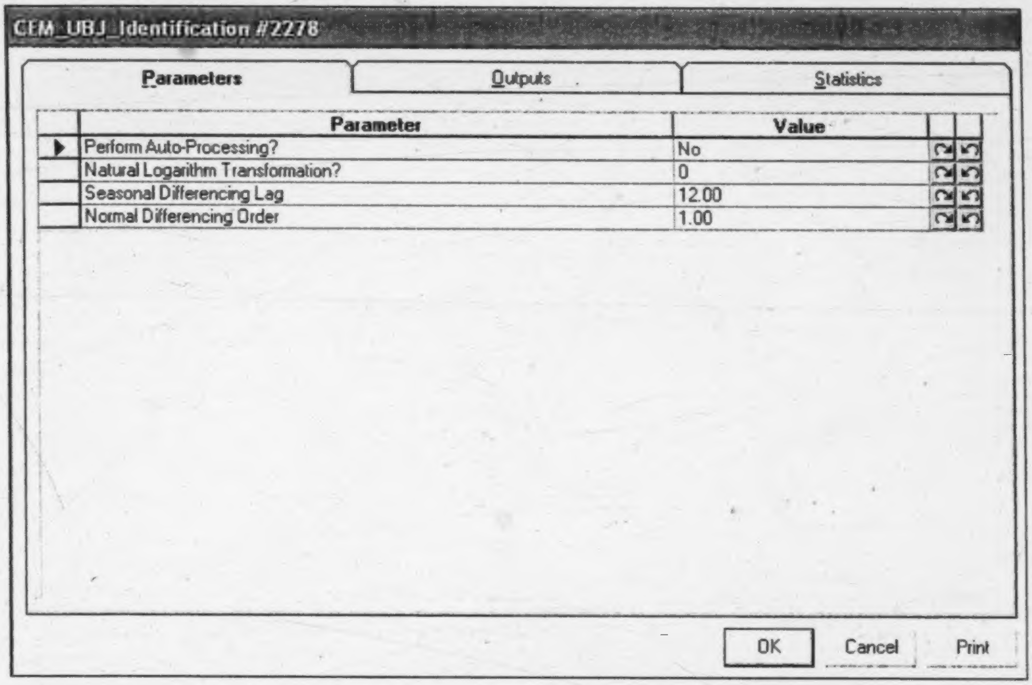

Data Processing.2278 Attribute: Partial Autoregressive Correlation Function (Processed ...
Partial Autoregressive Correlation Function (Processed Series)
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
-0.0
-0.1
-0.2
-0.3
-0.4
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20

Figure A-7: Screenshot of PACF Plot Window

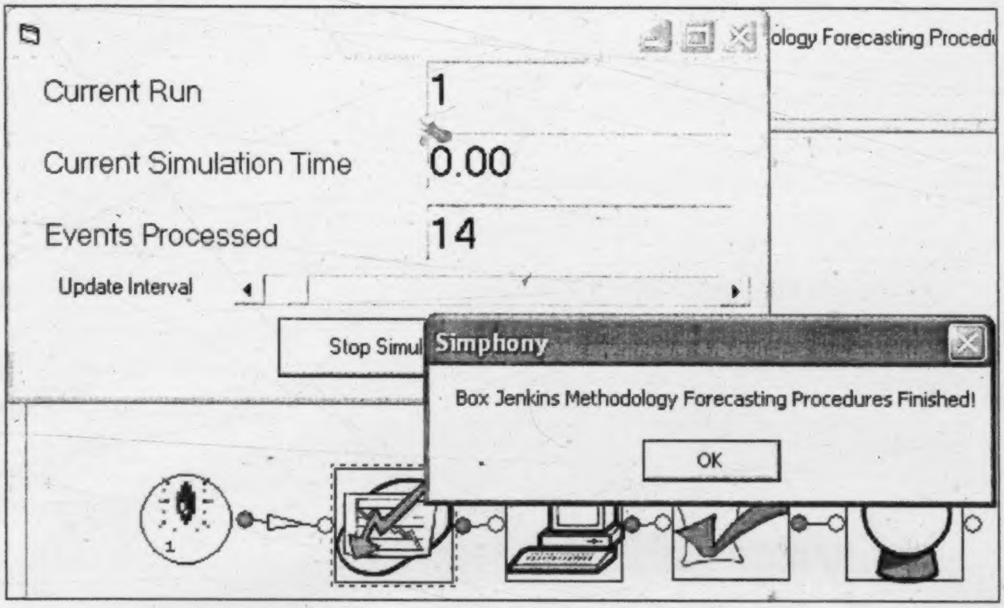
4) Modeling Parameters Updating


CEM_UBJ_Identification #2278
Parameters
Outputs
Statistics
OK
Cancel
Print

Parameter	Value
Perform Auto-Processing?	No
Natural Logarithm Transformation?	0
Seasonal Differencing Lag	12.00
Normal Differencing Order	1.00

Figure A-8: Screenshot of Modeling Parameters Updating Window

5) Forecasts output


ology Forecasting Proced
Update Interval
Stop Simul
Simphony
OK
1

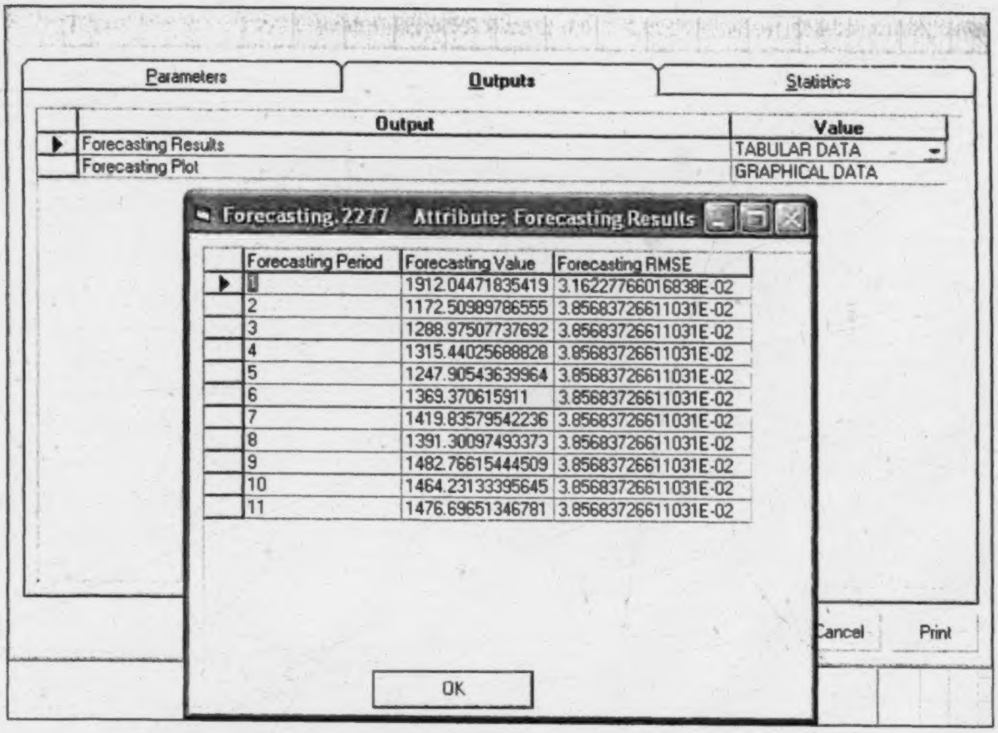
Current Run 1

Current Simulation Time 0.00

Events Processed 14

Box Jenkins Methodology Forecasting Procedures Finished!

Figure A-9: Screenshot of Modeling Termination Window


Parameters
Outputs
Statistics
Cancel
Print
OK

Output	Value
Forecasting Results	TABULAR DATA
Forecasting Plot	GRAPHICAL DATA

Forecasting.2277 Attribute: Forecasting Results

Forecasting Period	Forecasting Value	Forecasting RMSE
1	1912.04471835419	3.16227766016838E-02
2	1172.50989786555	3.85683726611031E-02
3	1288.97507737692	3.85683726611031E-02
4	1315.44025688828	3.85683726611031E-02
5	1247.90543639964	3.85683726611031E-02
6	1369.370615911	3.85683726611031E-02
7	1419.83579542236	3.85683726611031E-02
8	1391.30097493373	3.85683726611031E-02
9	1482.76615444509	3.85683726611031E-02
10	1464.23133395645	3.85683726611031E-02
11	1476.69651346781	3.85683726611031E-02

Figure A-10: Screenshot of Forecasts Output Window

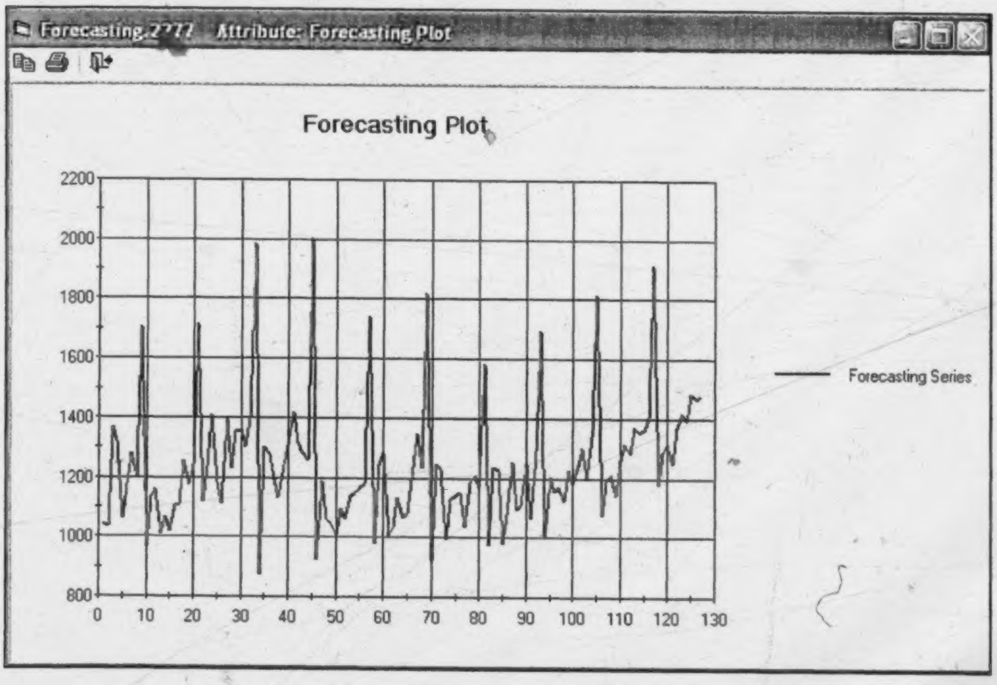

Forecasting.2277 Attribute: Forecasting Plot
Forecasting Plot
2200
2000
1800
1600
1400
1200
1000
800
0
10
20
30
40
50
60
70
80
90
100
110
120
130
Forecasting Series

Figure A-11: Screenshot of Forecasts in Time-series Plot Window

APPENDIX B: Box Jenkins-based Monte Carlo Simulation Validation

The following tables tabulate the data collection of all resulting simulated observations for the testing eleven periods using Box Jenkins-based Monte Carlo simulation. These sampled data are employed to perform basic simulation output analyses, which methodologies have been discussed briefly in Chapter 3. The time-series plot of these simulated observations is also provided in Chapter 3.

It is noted that the Box Jenkins-based Monte Carlo simulation is executed by 100 iterations for demonstration purpose. Theoretically, it is recommended that as many as possible iterations should be attempted to make the simulation calculation stably converge in order to mimic the real complex variation of the data being studied.

	Sep-02	Oct-02	Nov-02	Dec-02	Jan-03	Feb-03	Mar-03	Apr-03	May-03	Jun-03	Jul-03
Original	1901.00	1188.00	1354.00	1373.00	1276.00	1261.00	1279.00	1434.00	1466.00	1535.00	1449.00
Iteration 1	1658.69	1115.12	1262.03	1083.55	1318.93	1494.80	1413.24	1438.49	1521.76	1456.27	1598.97
Iteration 2	2244.13	1251.98	1400.26	1576.89	1405.08	1505.69	1439.71	1544.36	1401.31	1633.34	1816.90
Iteration 3	2042.56	1402.64	1143.50	1120.04	1292.43	1292.41	1490.30	1512.55	1625.68	1727.79	1704.92
Iteration 4	1729.14	1114.30	1209.03	1008.14	1150.11	1083.55	1461.48	1160.18	1417.58	1376.25	1225.19
Iteration 5	1903.04	1016.06	1333.33	1371.77	1507.75	1539.40	1429.17	1526.85	1424.42	1510.09	1517.29
Iteration 6	1863.69	1324.48	1039.59	1330.22	1353.45	1490.98	1551.71	1467.93	1658.11	1654.20	1567.35
Iteration 7	1867.04	934.54	1172.08	1241.42	1209.19	1516.80	1360.56	1464.10	1610.92	1650.64	1412.27
Iteration 8	1948.23	1030.14	1140.76	1410.67	1208.18	1402.92	1425.36	1290.82	1364.66	1473.04	1293.89
Iteration 9	1998.59	1073.64	1236.97	1110.97	834.59	1339.99	1206.81	1283.16	1375.45	1335.33	1192.80
Iteration 10	1855.09	974.94	1056.74	1419.33	1203.88	1134.26	1307.74	1111.95	1118.02	1203.07	1388.30
Iteration 11	2011.90	1025.01	1195.44	1230.90	1006.71	1056.32	1388.57	1262.84	1254.98	1427.52	1389.65
Iteration 12	2115.81	1278.04	1588.72	1631.21	1338.75	1336.36	1512.94	1441.86	1440.12	1586.90	1772.91
Iteration 13	1783.24	1096.17	1339.65	1165.77	1006.22	1237.04	1290.31	1342.52	1291.87	1313.81	1276.04
Iteration 14	2145.56	1150.73	1126.66	1367.62	1112.79	1335.68	1286.88	1406.43	1525.13	1416.89	1138.51
Iteration 15	2154.35	1400.54	1168.45	1350.80	1187.93	1225.95	1197.47	1343.96	1339.45	1375.00	1325.58
Iteration 16	1968.47	1300.32	1660.72	1256.39	1177.46	1565.41	1446.35	1249.11	1535.27	1719.49	1674.67
Iteration 17	1988.79	1479.29	1346.03	1235.22	1256.62	1581.56	1335.82	1522.47	1454.98	1584.16	1370.79
Iteration 18	1607.13	1005.49	1244.61	1173.17	1327.82	1221.93	985.75	1016.83	1399.68	1251.80	1333.98
Iteration 19	1589.50	977.11	1054.92	1001.31	1014.22	1386.38	1269.46	1431.71	1286.91	1343.47	1236.04
Iteration 20	1873.83	1234.01	1401.12	1344.71	1420.74	1799.04	1684.04	1597.99	1829.61	1722.60	1560.73
Iteration 21	1911.28	1130.26	1470.23	1644.57	1299.64	1492.34	1408.96	1569.71	1229.23	1612.44	1553.04
Iteration 22	2096.62	1186.43	1311.82	1159.16	973.51	1164.65	1332.56	1191.87	1137.97	1424.24	1214.46
Iteration 23	1911.92	1134.53	1337.81	1282.00	989.39	1131.07	1327.79	1082.30	1254.41	1129.56	1063.34
Iteration 24	1850.95	1180.49	1260.85	1461.51	1243.63	1160.32	1373.44	1354.73	1486.76	1510.64	1664.43
Iteration 25	1951.72	1182.48	1212.08	1625.13	1291.51	1225.93	1695.36	1528.34	1415.01	1609.69	1749.42
Iteration 26	1932.61	1157.57	1111.83	1441.92	1246.31	1450.98	1335.52	1283.36	1492.27	1456.59	1738.85
Iteration 27	2248.60	1503.54	1163.92	1141.34	1165.05	1319.16	1390.72	1488.55	1381.25	1329.68	1388.18
Iteration 28	1954.47	1346.29	1444.92	1139.07	1363.54	1534.38	1595.45	1698.61	1700.69	1545.04	1606.27
Iteration 29	1984.04	1212.02	1226.16	1215.81	1053.43	1282.41	1395.65	1276.57	1327.08	1356.17	1345.87
Iteration 30	2108.66	1298.91	1563.39	1378.10	1193.26	1556.39	1527.35	1379.41	1537.63	1222.99	1613.69
Iteration 31	1807.32	1344.66	1180.55	1240.79	1140.11	1561.71	1452.45	1250.88	1286.11	1651.03	1773.86
Iteration 32	1721.00	1082.38	1191.26	1285.63	1098.38	1131.92	1218.25	1401.77	1739.22	1503.56	1451.12
Iteration 33	1837.88	1433.43	1520.45	1242.63	1203.58	1600.16	1567.24	1557.28	1347.29	1544.75	1637.31
Iteration 34	1759.54	1098.25	1074.10	1201.63	1184.16	1484.87	1416.83	1286.75	1684.36	1617.79	1577.71
Iteration 35	1834.76	1279.45	1345.80	1438.82	1145.59	1611.89	1565.33	1199.04	1565.60	1691.00	1665.25
Iteration 36	2112.81	1094.76	957.11	1500.53	1302.46	1385.51	1372.07	1153.18	1441.33	1190.21	1220.99
Iteration 37	2027.03	1264.36	1136.27	1531.04	1311.10	1713.60	1381.51	1362.33	1420.20	1466.85	1342.08
Iteration 38	1550.15	935.80	1180.43	1151.60	1078.10	1199.57	1288.09	1165.42	1202.87	1336.51	1230.98
Iteration 39	1766.09	1106.25	1351.48	1163.41	1211.61	1282.39	1346.45	1271.39	1084.91	1265.20	1504.85
Iteration 40	1854.68	1054.59	1279.17	1220.86	1113.67	1299.69	1534.98	989.96	1438.04	1204.05	1467.21
Iteration 41	1878.00	1171.07	1216.78	1257.47	981.33	1092.17	1138.29	1252.08	1215.46	1493.03	1363.52
Iteration 42	1842.96	1052.47	1253.91	1235.84	1149.29	1188.80	1307.37	1066.04	1404.01	1119.12	1171.52
Iteration 43	1949.78	1172.24	1561.24	1649.82	1505.31	1574.20	1492.07	1479.14	1405.74	1479.56	1460.17
Iteration 44	2026.62	1216.97	1282.00	1352.16	1392.37	1772.71	1635.33	1703.40	1806.51	1800.21	1906.11
Iteration 45	1918.76	1126.75	1311.07	1465.06	1476.62	1386.40	1633.26	1481.66	1540.30	1639.82	1736.56
Iteration 46	1684.09	1096.06	1301.99	1282.43	1445.98	1349.59	1556.13	1323.77	1306.45	1523.48	1253.30
Iteration 47	1820.01	918.31	1277.83	1202.79	967.93	1091.65	1086.41	1151.76	883.55	1015.84	890.41
Iteration 48	2003.09	1222.37	1307.78	944.20	1160.79	1145.98	1249.14	1198.97	1149.04	1341.76	1329.08
Iteration 49	2047.74	1453.54	1233.81	1374.03	1140.31	1493.94	1540.74	1380.40	1441.74	1368.01	1601.31
Iteration 50	1889.25	974.21	1271.88	1674.37	1365.17	1239.44	1478.56	1205.42	1471.97	1180.12	1481.80

Table B-1: Simulated Observations for Testing Period of Time (Iteration 1-50)

	Sep-02	Oct-02	Nov-02	Dec-02	Jan-03	Feb-03	Mar-03	Apr-03	May-03	Jun-03	Jul-03
Iteration 51	2148.11	1454.21	1308.41	1482.73	1506.77	1368.09	1563.32	1629.39	1443.89	1417.21	1649.26
Iteration 52	1991.83	1404.11	1449.00	1304.55	1434.99	1451.20	1395.24	1771.92	1653.17	1483.55	1692.27
Iteration 53	2002.34	1319.95	1368.51	1359.67	1441.45	1440.80	1523.43	1406.77	1449.90	1463.34	1323.55
Iteration 54	1927.63	1170.98	949.91	1137.39	1162.02	1313.17	1351.68	1298.45	1149.99	1353.36	1309.84
Iteration 55	2003.12	1186.91	1361.99	1302.93	1221.94	1319.49	1265.55	1373.80	1628.35	1390.62	1388.24
Iteration 56	1918.05	1408.77	1102.79	1152.68	1078.76	1612.40	1497.36	1355.49	1498.65	1428.91	1623.64
Iteration 57	2036.47	1316.89	1323.58	1357.32	1321.32	1307.60	1406.67	1389.87	1335.19	1567.94	1308.68
Iteration 58	1734.15	1280.89	1180.70	1016.92	1256.37	1510.80	1510.38	1431.54	1377.53	1335.86	1615.26
Iteration 59	1819.56	1267.23	1331.10	1277.69	1123.26	1376.73	1120.06	1045.42	1610.56	1367.00	1787.00
Iteration 60	1938.40	1320.89	1312.15	1278.22	1483.91	1320.94	1586.65	1394.68	1653.47	1684.94	1481.12
Iteration 61	2069.15	1059.90	1343.18	1325.59	1075.22	1285.23	1340.50	1164.19	1073.48	1124.15	1232.36
Iteration 62	2104.17	1139.16	1005.45	1267.99	1414.96	1386.49	1455.96	1406.14	1526.86	1624.08	1570.58
Iteration 63	1989.32	1215.89	1360.96	1208.17	1101.42	1321.16	1426.92	1224.65	1534.67	1446.55	1862.89
Iteration 64	1713.26	1170.90	1322.58	1218.30	1240.75	1246.33	1363.45	1396.51	1624.61	1596.68	1125.22
Iteration 65	2037.70	1209.19	1457.10	1510.25	1559.07	1729.16	1794.64	1412.62	1529.16	1720.05	1511.45
Iteration 66	2016.51	1207.09	1483.24	1192.73	1340.56	1162.56	1167.12	1336.49	1375.40	1199.84	1287.86
Iteration 67	1775.70	1169.18	1032.65	1265.01	856.14	1273.15	1319.68	1370.00	1275.69	1434.10	1211.18
Iteration 68	1589.10	1118.36	1093.04	1079.75	1117.45	1171.55	1405.46	1257.51	1514.53	1316.04	1389.48
Iteration 69	2081.85	1318.13	1324.94	1173.30	1147.13	1391.89	1255.36	1377.12	1606.04	1507.18	1518.72
Iteration 70	2097.01	1448.65	1255.35	1421.93	1422.16	1440.92	1898.99	1735.63	1606.75	1802.09	1834.84
Iteration 71	1651.49	1327.68	1036.03	1402.24	1048.07	1346.03	1502.37	1495.80	1505.99	1342.66	1436.45
Iteration 72	1658.78	958.97	1260.66	1308.93	1119.39	1294.35	1407.83	1360.99	1470.49	1741.11	1186.19
Iteration 73	1956.14	1286.55	1244.33	1467.52	1331.84	1564.17	1671.08	1741.27	1682.55	1871.18	1778.45
Iteration 74	1889.27	1088.83	995.90	1142.51	1025.77	1322.19	1347.73	1417.44	1458.38	1387.51	1401.80
Iteration 75	2089.23	1348.54	1173.22	1362.06	1247.89	1479.91	1424.00	1599.88	1543.60	1320.95	1701.48
Iteration 76	1998.05	1092.96	1409.93	1007.50	1344.56	1349.20	1154.24	1110.28	1363.60	1356.84	1192.81
Iteration 77	2124.92	1166.48	1336.99	1481.48	1465.52	1512.71	1302.24	1195.21	1526.77	1350.56	1527.26
Iteration 78	1982.06	1289.37	1396.77	1228.76	1216.03	1441.59	1203.23	1236.90	1238.94	1487.31	1656.40
Iteration 79	1798.28	1172.99	1231.15	1151.56	1227.23	1328.68	1195.14	1183.03	1304.78	1415.36	1168.39
Iteration 80	1838.76	1299.98	1580.04	1636.47	1556.03	1630.37	1656.71	1392.65	1579.79	1602.47	1651.92
Iteration 81	1810.95	1226.91	1103.30	1138.04	1194.09	1520.26	1247.63	1426.01	1329.71	1348.20	1302.36
Iteration 82	2042.63	891.68	1176.84	1325.84	962.77	1218.10	1225.78	1317.55	1224.10	1310.21	1234.85
Iteration 83	1853.59	1085.21	1205.98	1217.58	940.40	1397.04	1532.70	1344.29	1337.06	1429.32	1358.16
Iteration 84	1644.63	1348.39	1335.51	1436.11	1535.70	1173.31	1329.95	1407.19	1560.24	1401.50	1518.09
Iteration 85	1883.19	1117.91	1411.98	1255.68	1107.73	1189.13	1376.79	1205.20	1619.78	1518.30	1518.55
Iteration 86	1960.08	1391.40	1024.75	1485.56	1223.14	1385.18	1569.94	1536.05	1320.11	1244.30	1503.71
Iteration 87	1663.91	1043.00	1137.13	1029.17	916.16	1199.71	928.74	1303.08	1356.11	1225.06	1263.24
Iteration 88	1753.29	1181.42	1395.35	1520.54	1297.68	1526.00	1808.99	1917.39	1711.27	2072.88	1823.38
Iteration 89	1997.76	1190.72	1489.53	1492.62	1119.33	944.66	1399.07	1159.65	1496.40	1622.11	1317.59
Iteration 90	1936.15	1090.51	1508.74	1289.44	1463.54	1695.53	1260.02	1902.01	1886.49	1883.65	1667.90
Iteration 91	1703.37	916.71	1031.68	1242.80	1392.99	1523.79	1356.37	1048.78	1407.57	1252.00	1071.90
Iteration 92	2000.45	1329.49	1314.47	1305.72	1251.22	1368.02	1470.10	1674.65	1605.42	1673.16	1932.92
Iteration 93	2023.10	1250.74	1206.82	938.04	777.76	1099.22	1093.53	1067.79	987.28	1153.64	1173.62
Iteration 94	2027.31	957.35	1087.58	1188.01	949.69	1396.66	1543.49	1392.10	1244.89	1475.72	1029.32
Iteration 95	2399.18	1386.32	1281.23	1564.71	1359.89	1668.94	1553.94	1768.29	1870.68	1777.53	1721.76
Iteration 96	1883.02	1151.72	1383.04	1214.01	1109.19	1314.61	1255.32	1077.20	1459.40	1308.01	1110.38
Iteration 97	1890.16	810.84	968.73	863.19	1157.09	1139.22	1374.04	1312.39	1417.23	1244.54	1372.51
Iteration 98	1801.92	1038.34	1255.47	987.84	899.85	1198.78	1451.74	1322.90	1308.09	1329.19	1358.64
Iteration 99	1900.96	1311.97	1316.62	1229.70	1294.30	1542.90	1228.51	1210.80	1388.95	1461.27	1283.95
Iteration 100	2006.79	998.02	1415.65	1462.16	1205.22	1495.31	1672.52	1554.39	1743.20	1652.94	1654.26

Table B-2: Simulated Observations for Testing Period of Time (Iteration 51-100)

APPENDIX C: Developed Computing Codes for Box Jenkins Forecasting and Box Jenkins-based Monte Carlo Simulation

The developed computer codes for the implementation of Box Jenkins forecasting and Box Jenkins-based Monte Carlo simulation are written in the format of *MATLAB* M files. The resulting computing algorithms can be executed in *MATLAB* execution environment and eventually be compiled into standalone (Dynamic Link Library) DLL files. The resulting DLL files can be employed by other computer programs, which are coded by other programming languages, such as *Visual Basic*, C++, through the execution of specially designed computer codes.

The following sections will list the developed *MATLAB* codes of most related algorithms for Box Jenkins forecasting and Box Jenkins-based Monte Carlo simulation. Certain brief explanations will be provided to facilitate the comprehension and all the codes have been tested and executed in *MATLAB* 13.0 environment.

- **Natural Logarithm Transformation**

This algorithm is simple to be implemented. Therefore corresponding *MATLAB* codes are not developed.

- **Normal Differencing**

```
Function y=differencing (x, n)
% This function performs N order differencing transformation on input matrix.
% Input:
%      x: input matrix
%      n: Order of differencing
% Output:
%      y: differenced matrix

[rows , columns]  =  size(x);
```

```
if (rows ~= 1) & (columns ~= 1)
   error(' Input "Series" must be a vector.');
end

y=diff(x,n,1);
```

- **Seasonal Differencing**

```
Function [NonSeasonalSeries]=seasonaldiff(Series,lag)
% This function performs seasonal differencing transformation on input matrix.
% Input:
%      Series: input matrix with m row and 1 column
%      Lag: seasonal differencing lag, for example 4,6,8,12,24
% Output:
%      NonSeasonalSeries: output matrix with m-p row and 1 column

[r,c]=size(Series);
if lag>0 & lag<r
   for i=1:r-lag
      NonSeasonalSeries(i,c)=Series(i+lag)-Series(i);
   end
else
   error('Invalid Lag!');
   NonSeasonalSeries=Series;
end
```

- **ACF Calculation**

```
Function [ACF, Lags, Bounds] = ACF (Series)
% This function calculates ACF for input series
% Input:
%      Series: input series
% Output:
%      ACF: ACF for input series
%      Lags: ACF Lags
%      Bounds: ACF bounds

[ACF, Lags, Bounds]=autocorr (Series);
```

```
Lags    = Lags (2:length(Lags));
ACF=ACF (2:length(ACF));
```

- **PACF Calculation**

```
Function [PACF, Lags, bounds] = PACF(Series)
% This function calculates PACF for input series
% Input:
%       Series: input series
% Output:
%       PACF: PACF for input series
%       Lags: PACF Lags
%       bounds: PACF bounds

[partialACF , Lags , bounds]=parcorr(Series);
PACF=partialACF(2:length(partialACF));
Lags    = Lags(2:length(Lags));
```

- **T test**

```
Function [h,stats,size] = ttest(x,n,flag)
% This function calculates t-test statistics for input matrix x and
% performs t-test hypothesis test.
% Input:
%       x: input matrix
%       n: total number of original series data
%       flag:0 for ACF and 1 for PACF
% Output:
%       h: binary value for the result of t-test. 0 pass, 1 not pass.
%       stats: t-test statistics
%       size: total number of t-test statistics

if nargin < 1,
  error('Requires at least one input argument.');
end

[m1, n1] = size(x);
if (m1 ~= 1 & n1 ~= 1),
    error('First argument has to be a vector.');
end
x = x(~isnan(x));
size = length(x);
if flag==0
    rootsum=zeros(size,1);
    for i=2:size
        for j=1:i-1
            rootsum(i)=rootsum(i)+x(j)^2;
        end
    end
    stats = x.*sqrt(n)./sqrt(1+2.*rootsum);
else
    stats=x.*sqrt(n);
end
% Rule of Thumb for critical level of t-test
crit = 2;

% Determine if the actual significance exceeds the desired significance
h = 0;
n=0;
for i=2:size
    if abs(stats(i))>=crit
        n=n+1;
    end
end
if n>0
    h=1;
end
```

- **Q test**

```
Function [H, pValue, Qstatistic, criticalValue] =qtest(Series)
% This function perform Q test for input series
% Input:
%       Series: input series
% Output:
%       H: binary value for Q test. 1 pass and 0 not pass
%       pValue: significance levels
%       Qstatistic: Q test statistics
%       criticalValue: critical value

[H, pValue, Qstatistic, criticalValue] = lbqtest(Series);
if H==0
   H=1;
else
   H=0;
end
```

- **Parameters Estimation**

```
Function[Parameters,SEE,Innovations,ISCondition,Converge]=estimating(ROrder,MOrder,S
eries)

% This function calculate the parameters for ARIMA models and output the
% parameter sets, warning message and other useful information
% Input:
%       ROrder: order of AR model
%       MOrder: order of MA model
%       Series: input series
% Output:
%       Parameters: fitted ARIMA model parameters
%       SEE: squared estimation error
%       Innovations: estimation residuals
%       ISCondition: binary value for Stationarity/Invertiblity conditions.
%       0 satisfied 1 not satisfied
%       Converge: binary value for converge condition. 1 converge 0 not
%       converge

if ROrder==0
   InitAR=[];
else
```

```
  for i=1:ROrder
    InitAR(i)=0.1;
  end
end
if MOrder==0
  InitMA=[];
else
  for i=1:MOrder
    InitMA(i)=0.1;
  end
end
[r,c]=size(Series);
InitConstant=mean(Series,1)*(1-0.1*ROrder);

spec=garchset('R',ROrder,'M',MOrder,'C',InitConstant,'AR',InitAR,'MA',InitMA,'Display','Of
f');
[Coeff,Errors,LLF,Innovations,Sigma,Summary]=garchfit(spec,Series);
if ROrder==0
  for i=1:3
    RPara(i)=0;
  end
else
  RPara=garchget(Coeff,'AR');
  for i=ROrder+1:3
    RPara(i)=0;
  end
end

if MOrder==0
  for i=1:3
    MPara(i)=0;
  end
else
  MPara=garchget(Coeff,'MA');
  for i=MOrder+1:3
    MPara(i)=0;
  end
end
Constant=garchget(Coeff,'C');
Parameters=[Constant,RPara,MPara];
Converge=0;
if Summary.converge=='Function Converged to a Solution'
  Converge=1;
end
ISCondition=0;
if Summary.warning=='No Warnings'
  ISCondition=1;
end

if ROrder==0
  for i=1:3
    RParaSEE(i)=0;
  end
else
RParaSEE=Errors.AR;
  for i=ROrder+1:3
    RParaSEE(i)=0;
  end
end
if MOrder==0
  for i=1:3
    MParaSEE(i)=0;
  end
else
  MParaSEE=Errors.MA;
  for i=MOrder+1:3
    MParaSEE(i)=0;
  end
```

```
end
ConstantSEE=Errors.C;
SEE=[ConstantSEE,RParaSEE,MParaSEE];
```

- **Forecasting**

```
Function[MeanForecast,MeanRMSE]=forecasting(ROrder,MOrder,Constant,AR,MA,Series,NumPeriods)
% This function forecast predetermined periods of future outcomes according
% to fitted ARIMA model
% Input:
%     ROrder: order of fitted AR model
%     MOrder: order of fitted MA model
%     Constant: fitted constant
%     AR: fitted AR model parameters
%     MA: fitted MA model parameters
%     Series: input series
%     NumPeriods: input periods of future time
% Output:
%     MeanForecast: mean forecast
%     MeanRMSE: mean RMSE

if ROrder==0
   AR=[];
end
if MOrder==0
   MA=[];
end
Spec=garchset('R',ROrder,'M',MOrder,'C',Constant,'AR',AR,'MA',MA,'K',0.001,'P',0,'Q',0,'Display','Off');
[SigmaForecast, MeanForecast, SigmaTotal, MeanRMSE] = garchpred(Spec, Series, NumPeriods);
```

- **Inverse Normal Differencing**

```
function [NewSeries]=invdifferencing(SeriesBeforeDiff,ForecastSeries,Order)

% This function transform the differenced series back to raw series
% Input:
%       SeriesBeforeDiff: series before normal differencing processing
%       ForecastSeries: forecasts series
%       Order: order of normal differencing
% Output:
%       NewSeries: Un-differenced series

[r1,c1]=size(SeriesBeforeDiff);
[r3,c3]=size(ForecastSeries);
NewSeries=SeriesBeforeDiff;
switch Order
  case 0
    NewSeries=[NewSeries;ForecastSeries];
  case 1
    for i=1:r3
      NewSeries(i+r1,c1)=ForecastSeries(i,c3)+NewSeries(i+r1-1,c1);
    end
  case 2
    for i=1:r3
      NewSeries(i+r1,c1)=ForecastSeries(i,c3)+2*NewSeries(i+r1-1,c1)-NewSeries(i+r1-
2,c1);
    end
  case 3
    for i=1:r3
      NewSeries(i+r1,c1)=ForecastSeries(i,c3)+3*NewSeries(i+r1-1,c1)-
3*NewSeries(i+r1-2,c1)+NewSeries(i+r1-3,c1);
    end
    otherwise
    error('Cannot process Series with differencing order more than 3!');
end
```

- **Inverse Seasonal Differencing**

```
Function[NewSeries]=invseasonaldiff(SeriesBeforeSeasonalDiff,NonSeasonalSeries,Lag)
% This function performs reverse seasonal differencing on input series
% Input:
```

```
%      SeriesBeforeSeasonalDiff: series before seasonal differencing
%      NonSeasonalSeries: seasonal differenced series
%      Lag: seasonal differencing lag
% Output:
%      NewSeries:reverse seasonal differenced series

[r1,c1]=size(SeriesBeforeSeasonalDiff);
[r2,c2]=size(NonSeasonalSeries);
if Lag==0
  NewSeries=NonSeasonalSeries;
else
  point=r1-Lag;
  NewSeries=SeriesBeforeSeasonalDiff;
  for i=1:r2-point
    NewSeries(r1+i,c1)=NewSeries(r1+i-Lag,c1)+NonSeasonalSeries(point+i,c2);
  end
end
```

- **Inverse Natural Logarithm Transformation**

This algorithm is easy to be implemented so no specific MATLAB codes are developed.

- **Box Jenkins based Monte Carlo Sampling**

```
Function [ySim]=Simulation(ROrder,MOrder,Series,Horizon,nPaths)

% This function perform a Monte Carlo Simulation for specified ARIMA models
% and output the simulated series paths.
% Input:
%      ROrder: fitted AR model order
%      MOrder: fitted MA model order
%      Series: input series
%      Horizon: Periods of future time
%      nPaths: Number of iterations
% Output:
%      ySim:simulated future outcomes in matrix

if ROrder==0
```

```
  InitAR=[];
else
  for i=1:ROrder
    InitAR(i)=0.1;
  end
end
if MOrder==0
  InitMA=[];
else
  for i=1:MOrder
    InitMA(i)=0.1;
  end
end
[r,c]=size(Series);
InitConstant=mean(Series,1)*(1-0.1*ROrder);

spec=garchset('R',ROrder,'M',MOrder,'C',InitConstant,'AR',InitAR,'MA',InitMA,'Display','O
ff');
[Coeff,Errors,LLF,eFit,sFit]=garchfit(spec,Series);
[eSim,sSim,ySim] = garchsim(Coeff,Horizon,nPaths,[],[],[],...
                    eFit,sFit,Series);
```

APPENDIX D: Box Jenkins Forecasting SPS Template: User Manual

D.1: Overview

The Box Jenkins forecasting SPS template is a tool for automating Box Jenkins forecasting modeling process and providing Box Jenkins-based Monte Carlo simulation input modeling. The template, in one aspect, allows average users to employ Box Jenkins forecasting method in an easy and timesaving approach, without damaging the powerful features of Box Jenkins methodologies in providing accurate short and medium range forecasts for complex situations. In another aspect, it provides a convenient tool to perform time series data fitting and parameters estimation, which are essential to Box Jenkins-based Monte Carlo simulation.

This template was developed in *Simphonoy* with SPS template development language. The structure of this template follows the basic format of a traditional SPS template, that is, one father element on the top and several child elements underneath. Basically, the father element executes the function of collecting inputs and storing global variables, while the child elements, collectively perform the functionalities, which Box Jenkins forecasting modeling requires.

The template provides the user with one father element and four child elements: (1) UBJ Forecasting Root, which is the highest element in the hierarchy of elements, (2) Data Processing, which represent the data processing functionality, (3) Parameter Estimating, which represent the parameter estimation functionality, (4) Diagnostic Check, which represent the diagnostic check functionality, and (5) Forecasting, which represent the calculation of forecasts functionality.

D.2: Procedures of creating a Box Jenkins forecasting Model

To create a new Box Jenkins forecasting model using the developed SPS template, the following steps can be followed:

1. Create a UBJ Forecasting Root;
2. Define the input parameters of the UBJ Forecasting Root, that is, input the historical data series and specify the forecasting parameters;
3. Create a New Entity element from Common template in child window as the source of new entity through the whole simulation model;
4. Create and link all four child elements after the New Entity element by the sequence of: Data Processing element-> Parameters Estimating element->Diagnostic Check element-> Forecasting element;
5. Define the input parameters for Data Processing element or leave them as default;
6. Link all the five elements in child window by collecting the connecting points from left to right;
7. Check the proper connection among elements and start the simulation;
8. Check the pop-up window for initial processing information or processing alert and take proper actions to refine the inputs for Data Processing element; and
9. Repeat Step 8 until successful completion information is displayed.

The following sections describe the details of using each element in the template.

D.3: Box Jenkins Forecasting Element

UBJ Forecasting Root

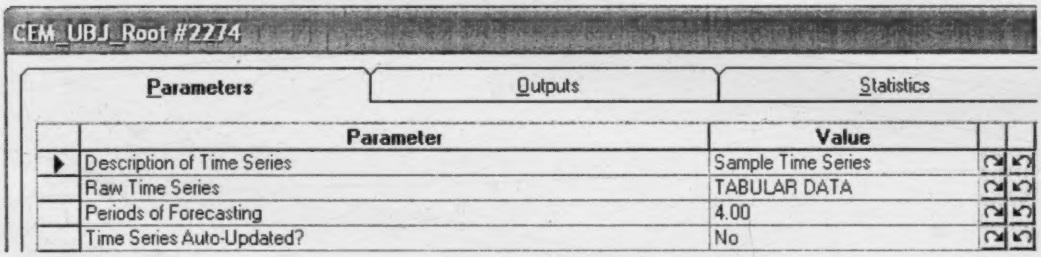
This element allows the initialization of Box Jenkins forecasting modeling. Users can input the original time series data, specify the desired periods of forecasting, and choose the method of data updating.

It is necessary to know that original time series data should be in the form of n rows time 2 columns. Here n refers to the number of data points, which is unlimited in theory. The first column is for the time period information and the second column is for the data related to the period.

Because Box Jenkins forecasting model is sensitive to the change of time series data, it is crucial to consider the updating of raw data when new data points are available. This template allows the flexibility of switching between manual data set updating and automatic data set updating. For automatic data set updating, it is assumed that the newest data point from the forecasting model is equal to the real data so it can be added to the end of raw data set for data updating.

Input:

Description of Time Series: text description for the raw data
Raw Time Series: original time series data for analysis
Periods of Forecasting: number of period of forecasting
Time Series Auto-Updated: Yes/No switch for data updating methods



	Parameter	Value
▶	Description of Time Series	Sample Time Series
	Raw Time Series	TABULAR DATA
	Periods of Forecasting	4.00
	Time Series Auto-Updated?	No

Figure D-1: Inputs Windows for UBJ Root Element

Outputs:

Forecasting Results: the calculated forecasting results for the periods of forecasting specified in the input windows.

Statistics:

None.

Data Processing

Data processing element performs the basic data transformation and processing operations for analysis purpose. These operations include normal differencing, seasonal differencing and natural logarithm transformation. The final product by these operations is a stationary time series data for next procedure of modeling.

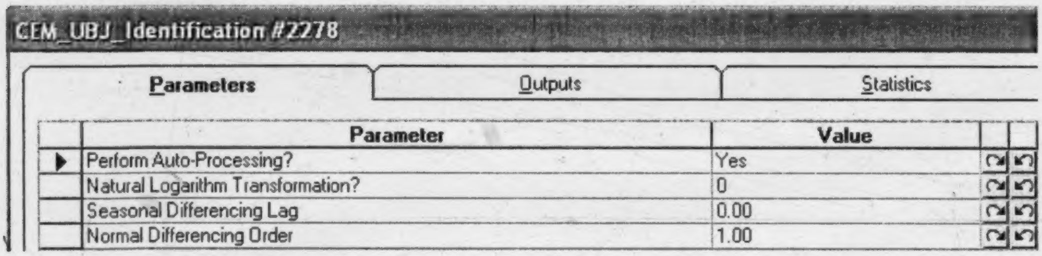
Inputs:

Perform Auto-Processing: Yes/No switch between hand modeling and automatic modeling.

Natural Logarithm Transformation: Yes/No switch for Natural Logarithm Transformation, only valid for hand modeling.

Seasonal Differencing Lag: Lag for seasonal differencing Lag, only valid for hand modeling.

Normal Differencing Order: Order for normal differencing, min 0 and max 3. only valid for hand modeling.



Parameter	Value
Perform Auto-Processing?	Yes
Natural Logarithm Transformation?	0
Seasonal Differencing Lag	0.00
Normal Differencing Order	1.00

Figure D-2: Inputs Windows for Data Processing Element

Outputs:

Plot for Raw Series: two dimensional x y plot for raw data.

Plot for Processed Series: two dimensional x y plot for processed data.

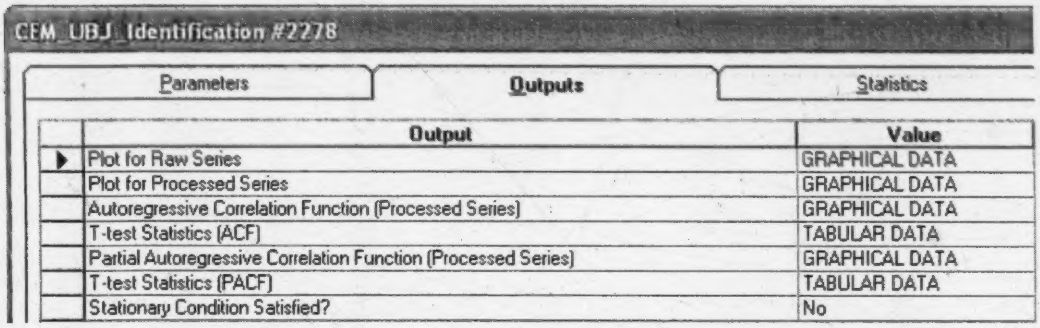
Autoregressive Correlation Function (Processed Series): ACF for processed data, represented by column plot.

T-test Statistics (ACF): T-test statistics values for ACF.

Partial Autoregressive Correlation Function (Processed Series): PACF for processed data.

T-test Statistics (PACF): T-test statistics values for PACF.

Stationary Condition Satisfied: Yes/No switch showing the stationary condition status.



Output	Value
Plot for Raw Series	GRAPHICAL DATA
Plot for Processed Series	GRAPHICAL DATA
Autoregressive Correlation Function (Processed Series)	GRAPHICAL DATA
T-test Statistics (ACF)	TABULAR DATA
Partial Autoregressive Correlation Function (Processed Series)	GRAPHICAL DATA
T-test Statistics (PACF)	TABULAR DATA
Stationary Condition Satisfied?	No

Figure D-3: Outputs Windows For Data Processing Element

Statistics:

None.

Parameters Estimation

This element performs the estimation calculation of model parameters for each tentative ARIMA model. It uses MLE (Maximum Likelihood Estimation) method to calculate the constant, parameters for each ARIMA components. The estimation errors are also provided for reference.

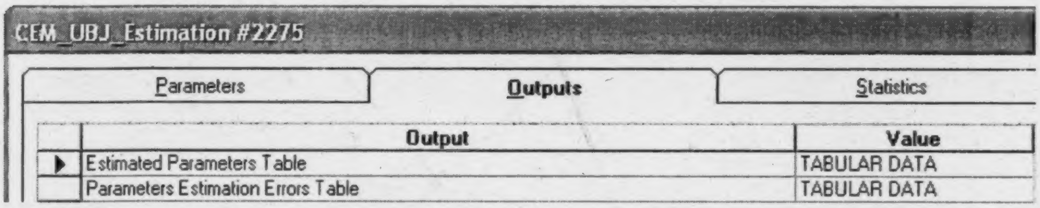
Inputs:

None.

Outputs:

Estimated Parameters Table: calculated model parameters and model specification.

Parameters Estimation Errors Table: calculated model parameters estimation errors.



Output	Value
Estimated Parameters Table	TABULAR DATA
Parameters Estimation Errors Table	TABULAR DATA

Figure D-4: Outputs Windows For Parameters Estimation Element

Statistics:

None.

Diagnostic Check

This element performs the completed diagnostic checks. An array table displays the checks results, in which zero value represents Failure and one represents Pass.

All the good models from tentative models will be identified but the Best ARIMA model will be chosen for real forecasting.

Inputs:

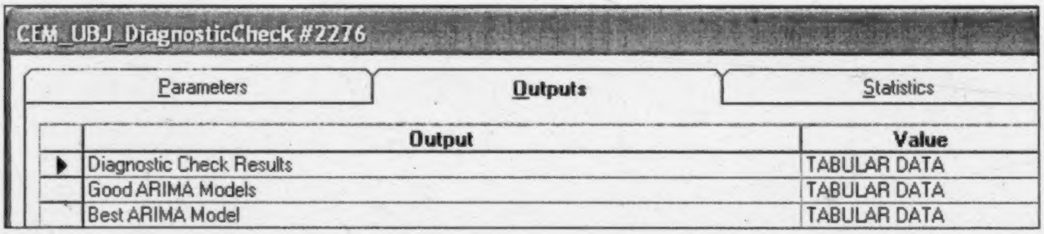
None.

Outputs:

Diagnostic Check Results: Array table shows the check results. Note: 0 Failure, 1 Pass.

Good ARIMA Models: Array table shows the ARIMA models who pass all the check items.

Best ARIMA Models: best ARIMA model with the least RMSE compared with other good models.



	Output	Value
▶	Diagnostic Check Results	TABULAR DATA
	Good ARIMA Models	TABULAR DATA
	Best ARIMA Model	TABULAR DATA

Figure D-5: Outputs Windows for Diagnostic Checks Element

Statistics:

None.

Forecasting

This element uses the best model identified by Diagnostic Check element to perform forecasting operation. Period of forecasting will be read from parent element and forecasting values will be displayed as an array table. A two-dimensional x y plot will be display to extend original time series to new periods.

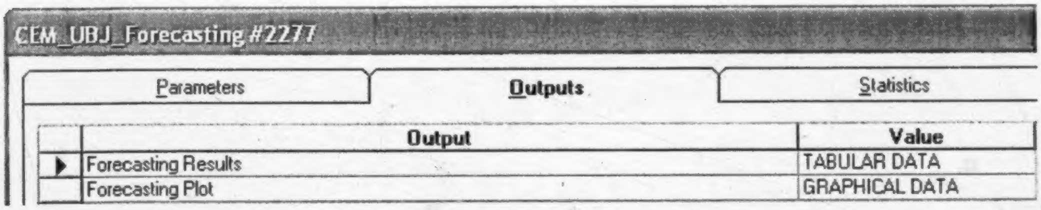
Inputs:

None.

Outputs:

Forecasting Results: Array table to show the forecasting results. These values will be sent up to the parent element too.

Forecasting Plot: two-dimensional x y plot to show the new time series with forecasting values.



Output	Value
Forecasting Results	TABULAR DATA
Forecasting Plot	GRAPHICAL DATA

Figure D-6: Outputs Windows for Forecasting Element

Statistics:

None.